COX'S PROPORTIONAL HAZARDS MODEL APPLIED TO LOAN BOOK DATA.

JOSH STRONG

Abstract. Survival Analysis is a branch of statistical mathematics that provides powerful tools for investigation into data for which there exists an event of interest. The essence of Survival Analysis is a specialised form of regression modelling. In this project, a brief introduction to Survival Analysis is given. Firstly, the foundations of Survival Analysis are explored. A popular regression model concerning this type of data, called Cox's Proportional Hazards model, is derived and explained. Methods for diagnosing an important assumption about this regression model as well as determination of the significance of predictor variables are discussed. A technique for dealing with circumstances in which this important assumption is violated is explored. This knowledge is then applied to peer-to-peer loan book data, where the impact of a loan's assigned grade on its probability of default is investigated. The results of this investigation were puzzling and quite unexpected, with the findings contradicting one another.

Contents

List of Figures

List of Tables

1. Introduction

Survival analysis is a group of statistical techniques used to investigate and model the time until an event of interest occurs in relation to one or more factors. It provides answers to questions such as: How do specific characteristics affect the rate at which the event of interest occurs? What proportion of a population will survive until a given time? How does the rate of survival between multiple groups compare?
In this project there will be a focus on a specific regression model, in the set of Survival Analysis mathematics, called Cox's Proportional Hazards model. This semi-parametric model has many applications, but is primarily used by medical professionals to analyse the association between relevant medical characteristics of their patients (such as blood pressure, age, weight, etc.) and their survival time. Some recent examples of applications of the model include an analysis of unemployment duration in Slovenia, by Bori and Kavkler[1]; Ihwah, used the model to analyse factors which could influence purchase decisions on products[2]; and Ni, who analysed the stock exchange market using the model[3].
The Proportional-Hazards model was developed by Sir David Roxbee Cox, a British statistician[4]. This paper has since become one of the most cited statistical papers, with over 46,000 citations as of November 2017 just from Google Scholar. Cox won the Kettering prize and Gold medal in 1990, for his outstanding contributions to the diagnosis and treatment of cancer through his development of the Proportional Hazard Regression model[5].
This project will be applying Cox's statistical model to peer-to-peer (P2P) loan book data, with the fundamental question being: how does a loan's assigned grade affect the probability of the loan defaulting?
Peer-to-peer (P2P) lending is a platform provided by online companies by which investors, looking to gain interest on their surplus cash, can lend money to borrowers, who are looking for extra cash to finance specific situations in their lives.
Funding Circle has become one of the most popular companies in the P2P lending business, quoting a 7% average annual return on investment[6]. P2P lending was first introduced in February 2005 by the company Zopa. Since then, Zopa has approved over 277,000 borrowers and lent over 2.82 billion to UK customers, of which 800 million was lent in the past 12 months alone[7].
In this project I will be using the loan book data provided by the P2P lending company LendingClub (LC), which is freely available at `https://www.lendingclub.com/info/download-data.action`. This data provides a vast amount of information on borrowers, including: loan status (current, completed or defaulted); loan length; grade of the loans (which is assigned by LC which varies according to risk of the loan defaulting); interest rate on the loan; loan purpose; housing situation of the borrower; employment length of the borrower; credit history of the borrower; and many more.

2. Introduction to Survival Analysis

Survival analysis involves two important functions: the *survival function* and the *hazard function*. Cox's model focuses on the hazard function, but it is important to have knowledge of the survival function in order to gain a wider knowledge of survival analysis in addition to understanding the hazard function itself. Kleinbaum and Klein give an intuitive definition of these two functions in their text Survival Analysis: A Self-Learning Text [8, pages 9-12].
Throughout this project, examples are given based on lung cancer data provided by the

TABLE 1. First 3 individuals of lung cancer time-to-event data.

Individual	Survival time	Censoring Status	Age	Sex
1	306	2	74	1
2	455	2	68	1
3	1010	1	56	1

`survival` package in R. I reproduced all the analysis in this project with all figures and tables being produced in `R`, for which the code can be seen at the end of the project.

2.1. **Censoring and Time-to-Event Data.**

Survival analysis is based upon *time-to-event data*. Such data has 3 key groups of information of individuals: *survival time, censoring status* and *covariates*.
Survival times give the time until the event-of-interest occurs for an individual. However, since studies cannot practically continue forever, some individuals do not experience the event-of-interest within the study. It is therefore natural to ask, what survival times are these individuals given? Censoring provides a solution to this problem and other problems such as individuals withdrawing from the study and becoming lost to follow-up[1].
Censoring status is a binary value, providing us an indication of whether or not an individual experienced the event-of-interest within the study or not. Typically[2], the censoring status for individuals who do experience this event-of-interest is set to 1, and their survival time is the observed survival time. For individuals who do not experience this event-of-interest in the study, their censoring status is typically set to 0 and their survival time is set to the time from when they entered the study up until the end of the study.
Covariates contain information about the individuals for which we might be interested in investigating relations between such covariates of individuals and their survival times.
The time-to-event data in this project will primarily concern a type of censoring called *right censoring*, in particular *type I censoring*. Type I censoring is the type of censoring that only considers an individual to have an observed survival time if it has survival time less than or equal to the fixed time of the study. All individuals with survival times greater than the length of the study are censored. There exists many other types of censoring: Klein and Moeschberger give a complete overview on the topic of censoring in their text *Survival Analysis: Techniques for Censored and Truncated data (2003)*[9, chapter 3]

As an example, consider the lung cancer data in which patients with advanced lung cancer are monitored. The event-of-interest here is death. The first 3 individuals are shown in Table 1. The survival time column displays the survival times for these individuals as previously described. The censoring status has 2 indicating the individual had experienced the event-of-interest and died; and 1 indicating the individual had been censored. The columns age and sex are the covariates of these individuals. Males are indicated by 1 and females by 2.

[1]Lost to follow-up individuals includes all individuals who were involved in the study at one point, but due to some reason the investigator is no longer able to reach this individual to gather data.

[2]Although normally 1 is for individuals who have experienced the event of interest and 0 otherwise, when creating a survival object required for formulating a Cox's PH regression model in `R`, different numbers for censoring status are used to indicate different types of censoring[25, Arguments: event].

2.2. **The Survival Function.**
The *survival function*, denoted $S(t)$, is interpreted as the probability of the an individual surviving past time t. Let T be the continuous random variable being the time until occurrence of this event of interest. The survival function is the complement of the cumulative distributive function of the random variable T:

$$S(t) = \Pr(T \geq t) = 1 - F(t) = \int_t^\infty f(x)\,dx. \tag{1}$$

We can graphically represent survival functions with *survival curves*, which give a useful representation to the probability of survival for an individual over time. When inspecting survival curves, it is worth noting that they display similar patterns: they all begin at 1, trend towards 0 as time passes and are strictly decreasing. The significant difference in the curves lies within the rate of change of each curve, which leads to our next survival quantifier, the *hazard function*.

2.3. **The Hazard Function.**
The *hazard function*, denoted $\lambda(t)$, is an alternative feature of the distribution of T. It calculates the instantaneous rate of the event of interest occurring at the specific time t. The function is given by:

$$\lambda(t) = \lim_{h\to 0} \frac{\Pr(t \leq T \leq t+h | T \geq t)}{h}.$$

This should be read as the probability that event of interest will occur in the time interval $[t, t+h)$, given that event has not happened before, divided by the length of the time interval. Taking the limit as $h \to 0$ obtains the relative risk rate of event T occurring. The hazard function is related to the survival function as demonstrated here:

$$\begin{aligned}
\lambda(t) &= \lim_{h\to 0} \frac{Pr(t \leq T \leq t+h \cap T \geq t)}{Pr(T \geq t)h} && \textit{(definition of conditional probability)} \\
&= \lim_{h\to 0} \frac{Pr(t \leq T \leq t+h)}{S(t)h} && \textit{(definition of the survival function)} \\
&= \frac{1}{S(t)} \cdot \lim_{h\to 0} \frac{Pr(T \leq t+h) - Pr(T \leq t)}{h} \\
&= \frac{1}{S(t)} \cdot \lim_{h\to 0} \frac{F(t+h) - F(t)}{h} = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} && \textit{(definition of the derivative).}
\end{aligned} \tag{2}$$

Unlike survival curves, plots of hazards functions do not follow a particular pattern, but instead only follow one condition: $h(t) \geq 0,\ \forall t \in [0, \infty)$.

2.4. **Kaplan-Meier estimate of the Survival function.**
A prominent method for non-parametic estimation of survival functions is the Kaplan-Meier[3] (KM) estimate. Published by Edward L. Kaplan and Paul Meier in the Journal of American Statistical Association (1958)[10], the KM estimate is a stepwise-decreasing function with constant survival times between successive observed (uncensored) event times. Klein and Moeschberger give a derivation of the estimator, which goes as follows [9, page 102]:

Let $0 = t_0 < t_1 < ... < t_k$ denote observed (uncensored) event times in a time-to-event data

[3]The Kaplan-Meier estimate is also known as the Product-Limit estimator in some literature.

set, with $S(t_0) = 1$ and $S(t_1), ..., S(t_k)$ denoting the probability of an individual surviving past time t_i. We can find $\hat{S}(t_i)$ through considering

$$S(t_i) = \frac{S(t_i)}{S(t_{i-1})} \times \frac{S(t_{i-1})}{S(t_{i-2})} \times ... \times \frac{S(t_1)}{S(t_0)} \times S(t_0).$$

We can estimate each fraction $\frac{S(t_i)}{S(t_{i-1})}, \frac{S(t_{i-1})}{S(t_{i-2})}, ...$ through

$$\begin{aligned}
\Pr(T > t_i | T \geq t_i) &= \frac{\Pr(T > t_i, T \geq t_i)}{\Pr(T \geq t_i)} && \textit{(Definition of conditional probability)} \\
&= \frac{\Pr(T > t_i)}{\Pr(T \geq t_i)} && \textit{(In a discrete setting)} \\
&= \frac{S(t_i)}{S(t_{i-1})}, && \textit{(Definition of the survival function (1))}
\end{aligned}$$

with an estimation for $\Pr(T > t_i | T \geq t_i)$ given by

$$\hat{\Pr}(T > t_i | T \geq t_i) = \frac{r_i - d_i}{r_i} = \frac{\text{Number of individuals alive at time } t_i}{\text{Number of individuals at risk at time } t_i}$$

where

- r_i denotes the number of individuals at risk of experiencing the event at time t_i;[4]
- d_i denotes the number of individuals who experience the event at time t_i.

The Kaplan-Meier estimate is therefore given by

$$\begin{aligned}
\hat{S}(t_i) &= \hat{Pr}(T > t_i | T \geq t_i) \times \hat{Pr}(T > t_{i-1} | T \geq t_{i-1}) \times ... \hat{Pr}(T > t_1 | T \geq t_1) \times 1. \\
&= \prod_{t_i \leq t} \left[\frac{r_i - d_i}{r_i} \right].
\end{aligned}$$

Greenwood's formula is an estimate to the Kaplan-Meier's variance. It is given by [9, page 92]

$$\hat{\text{Var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{r_i(r_i - d_i)}.$$

Therefore, a $(1-\alpha)\%$ confidence interval for each successive observed event time $t_1, ..., t_k$ is given by:

$$\left[\hat{S}(t_i) - Z_{(1-\frac{\alpha}{2})} \times e.s.e(\hat{S}(t_i)), \hat{S}(t_i) + Z_{(1-\frac{\alpha}{2})} \times e.s.e(\hat{S}(t_i)) \right]$$

where,

- $e.s.e.(\hat{S}(t)) = \sqrt{\hat{\text{Var}}(\hat{S}(t))}$;
- $Z_{(1-\frac{\alpha}{2})}$ denotes the $(1 - \frac{\alpha}{2})$ quantile of the standard Gaussian distribution.

Figure 1 is a plot for the estimated survival function including a risk table of the lung data using the Kaplan-Meier method. The covariate sex was fitted in the model and a survival curve with the corresponding 95% confidence intervals is provided for both males and females. The Kaplan-Meier estimate has the property that the estimated survival

[4] r_i can be equivalently considered as the number of individuals not yet experienced the event of interest just before time t_i.

TABLE 2. Kaplan-Meier survival time estimates, corresponding estimated variances and 95% confidence interval for the first 3 event times in lung cancer data.

t_i	r_i	d_i	$\hat{S}(t) = \prod_{t_i \leq t}\left[\frac{r_i - d_i}{r_i}\right]$	$\hat{\text{Var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{r_i(r_i - d_i)}$	95% C.I
11	103	1	$\frac{103-1}{103} \approx 0.990$	$0.990^2 \times (\frac{1}{103(103-1)}) \approx 9.3 \times 10^{-5}$	$[0.990, 0.990]$
12	102	1	$0.990 \times \frac{102-1}{102} \approx 0.981$	$0.981^2 \times (9.3 \times 10^{-5} + \frac{1}{102(102-1)}) \approx 1.83 \times 10^{-4}$	$[0.981, 0.981]$
13	101	1	$0.981 \times \frac{101-1}{101} \approx 0.971$	$0.971^2 \times (1.83 \times 10^{-4} + \frac{1}{101(101-1)}) \approx 2.66 \times 10^{-4}$	$[0.971, 0.970]$
⋮	⋮	⋮	⋮	⋮	⋮

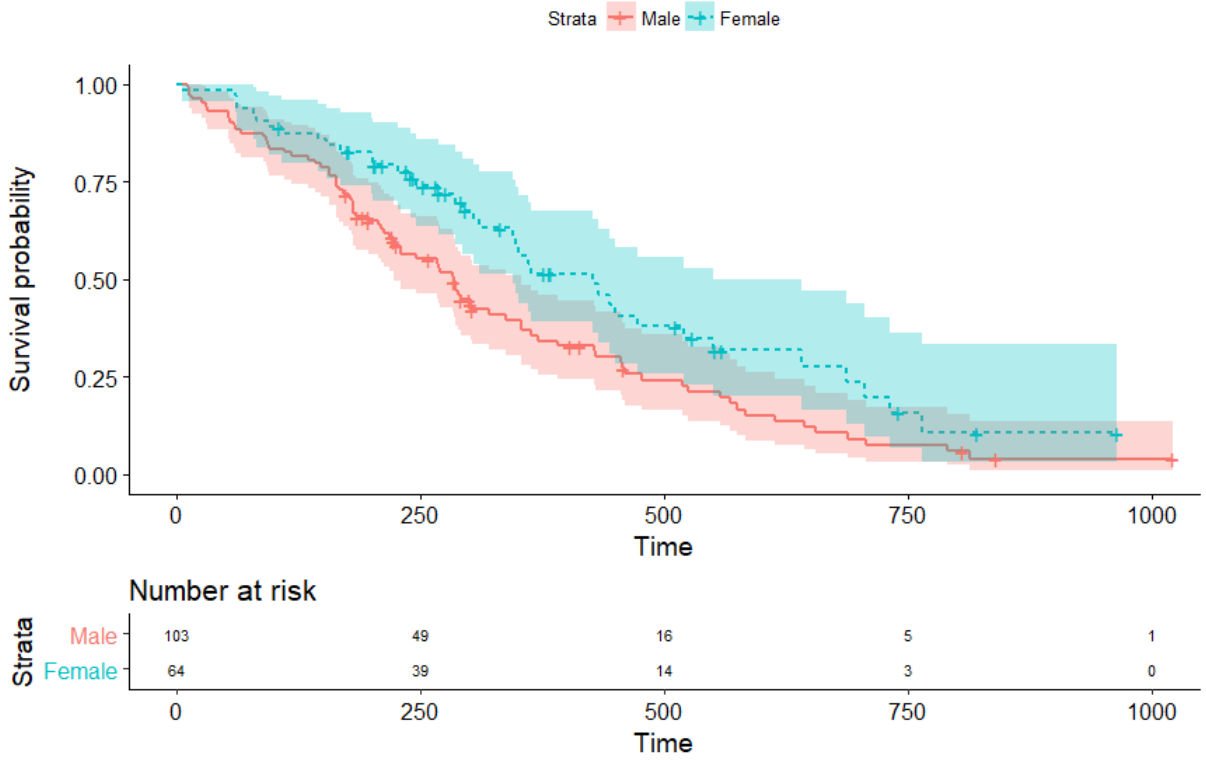


FIGURE 1. Survival curves and risk table for male and female patients with lung cancer.

probability for all times greater than the final event time is equal to that final event time i.e. $\hat{S}(t_k) \ \forall t \geq t_k$, which can be seen in at the tail ends of the survival curves in Figure 1. Table 2 shows the first 3 event times for males and their corresponding r_i, d_i, $\hat{S}(t_i)$, estimate variance and 95% confidence interval.

2.5. **Estimation of Mean and Median Survival time.**

Quantities such as the median and mean time to event of interest are useful pieces of information when dealing with survival data. Klein and Moeschberger derive the functions to calculate these quantities [9, pages 118-120]. We consider the latter first:
The *mean residual life function* $mrl(x)$ calculates the expected remaining time for an individual at time x until they experience the event of interest, given that they have survived up to this time. It is given by the total area under the survival function divided

by the probability of surviving past time x, as shown here:

$$\begin{aligned} mrl(x) &= \mathbb{E}\left[X - x | X > x\right] \\ &= \frac{\mathbb{E}[(X-x)I_{X>x}]}{\Pr(X>x)} && \textit{(Conditional expectation definition)} \\ &= \frac{\int_x^\infty (t-x)f(t)\,dt}{S(x)} && \textit{(Definition of expectation and survival function)} \\ &= \frac{-(t-x)S(t)|_x^\infty + \int_x^\infty S(t)\,dt}{S(x)} && \textit{(Evaluation of integral by parts)} \\ &= \frac{\int_x^\infty S(t)\,dt}{S(x)}. \end{aligned}$$

By setting $x = 0$ in the mean residual life function, we can obtain the expected time until an individual experiences the event of interest:

$$\begin{aligned} mrl(0) = \mu &= \frac{\int_0^\infty S(t)\,dt}{S(0)} \\ &= \int_0^\infty S(t)\,dt. && (S(0) = 1) \end{aligned} \tag{3}$$

Since in the Kaplan-Meier estimate for the survival function we have $\hat{S}(t) = \hat{S}(t_k)\ \ \forall t \geq t_k$, we have to set an upper limit τ to the integral in (3) otherwise the area under the survival function would be defined to be infinite. This limit is usually set to be the final observed event time $\tau = t_k$. An estimate for the expected time until an individual experiences the event time using the KM estimated survival function is therefore given by:

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)\,dt.$$

with estimated variance given by[9, page 118]

$$\hat{\text{Var}}(\hat{\mu}_\tau) = \sum_{i=1}^{k}\left[\int_{t_i}^{\tau} \hat{S(t)}\,dt\right]^2 \frac{d_i}{r_i(r_i - d_i)}$$

and hence a $(1-\alpha)\%$ confidence interval

$$\left[\hat{\mu}_\tau - Z_{1-\frac{\alpha}{2}}\sqrt{\hat{\text{Var}}(\hat{\mu}_\tau)}, \hat{\mu}_\tau + Z_{1-\frac{\alpha}{2}}\sqrt{\hat{\text{Var}}(\hat{\mu}_\tau)}\right]$$

where $Z_{1-\frac{\alpha}{2}}$ denotes the $(1-\frac{\alpha}{2})$ percentile of the standard Gaussian distribution.

The p^{th} quantile of the time of survival until the event of interest is given by

$$x_p = \inf\{t : S(t) \leq 1-p\}$$

with an estimation of this quantity, using the estimated survival function, is given by

$$\hat{x}_p = \inf\{t : \hat{S}(t) \leq 1-p\}.$$

The p^{th} quantile therefore satisfies

$$\hat{S}(\hat{x}_p) = 1 - p.$$

Following on from our lung data, using `R` we can easily obtain the estimated mean $\hat{\mu_\tau}$ and median (0.5^{th} quantile) $\hat{x}_{0.5}$ survival times for males and females (see Table 3). The average survival time of an male individual with lung cancer is seen to be 341.8 days, in comparison to the 465.5 days of a female individual. The median survival time for males is 286 and 426 for females. This data overall suggests that the prognosis for a female patient with lung cancer is better than for a male patient.

TABLE 3. Estimated mean and median survival times for male and females with from lung cancer data.

Covariate	μ_{t_k}	$\hat{x}_{0.5}$
Male	341.8	286
Female	465.5	426

3. Cox's Proportional Hazard Regression Model

3.1. **Cox's Hazard Function.**

The hazard function proposed by Cox[11], $\lambda(t)$, is given by:

$$\lambda(t) = \lambda_0(t) \exp(\sum_{i=1}^{p} \beta_j x_j)$$

where,

- t represents time;
- $\lambda_0(t)$ is the *baseline hazard*: the hazard function at time t corresponding to when all covariates are 0;
- x_j is the j^{th} covariate in the model;
- β_j is the coefficient for the j^{th} covariate in the model.

The β_j coefficients indicate the size of the effect of the covariates on the model. The *hazard ratios*, $\exp(\beta_j)$, illustrate association between covariates and event hazard:

- $\exp(\beta_j) > 1 \iff$ positive association with the event hazard.
- $\exp(\beta_j) = 1 \iff$ no association with the event hazard.
- $\exp(\beta_j) < 1 \iff$ negative association with the event hazard.

Following on from our lung data, we can see from Table 4 the estimated coefficients and hazard ratios for the model fitting the covariates sex and age. The hazard function for this model is:

$$\lambda(t) = \lambda_0(t) \exp\left(-0.45\text{sex} + 0.02\text{age}\right)$$

The covariate age is seen to have a slight positive association with the event hazard, meaning that older patients have a higher chance of dying. The covariate sex is a dichotomous variable, with 1 representing males and 2 representing females, therefore we can conclude that males have a higher chance of dying due to advanced lung cancer.

Cox's regression model is said to be a *semi-parametric* model, such that the baseline hazard is left completely unspecified. This model is the go-to regression model for Survival Analysis due to key underlying features.

Firstly, the main objective when dealing with such survival data is to find the unknown coefficients of the covariates, which give crucial information on the effect of each covariate on the time to the event of interest. Cox's regression allows for this to happen (under the proportional hazards assumption) without having to estimate the baseline hazard. This is seen in the following subsection.

Next, due to the nature of the exponential function being strictly positive on its entire domain ($\exp(x) > 0 \ \forall \ x \in \mathbb{R}$) it ensures the hazard ratios to be non-negative, which makes sense since the instantaneous rate of failure should never be a negative number.

Finally, usually we can not be completely sure on the best parametric distribution for modelling some data at hand. Cox's regression gives a good alternative with similar

TABLE 4. Estimated coefficients and hazard ratios of covariates sex and age.

Covariate	Estimated coefficient $\hat{\beta}$	Hazard ratio $e^{\hat{\beta}}$
Sex	-0.44669	0.63974
Age	0.01737	1.01752

results as to applying a parametric model to data. This can be seen in figure 2, where a parametric Weibull Accelerated Failure Time (AFT) model can be seen to give similar results to a Cox's PH regression model[5] when applied to the same data.

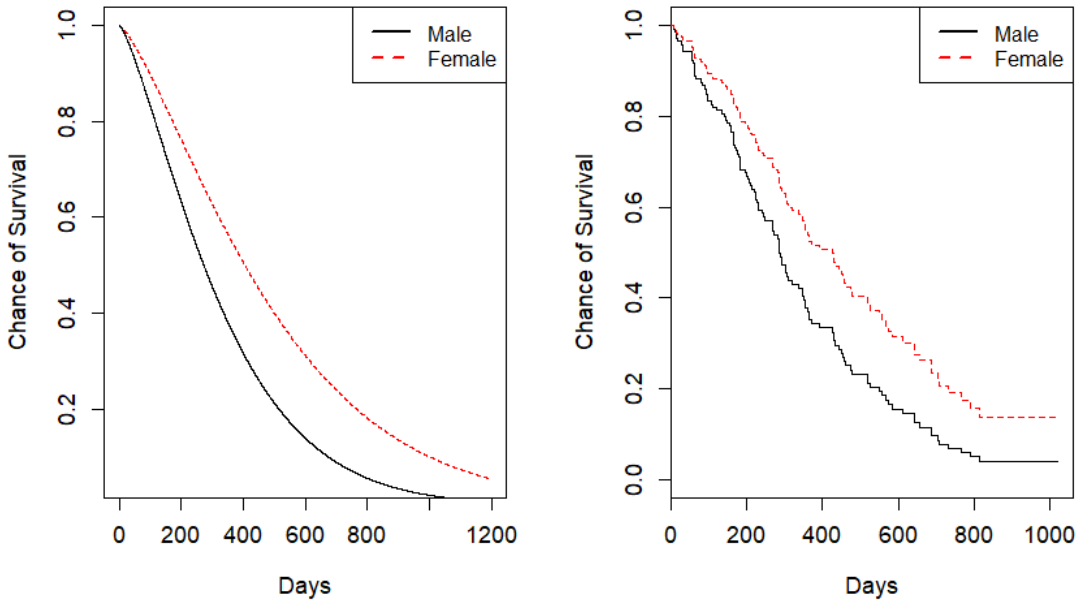


FIGURE 2. Comparison of estimated survival curves. First figure : Weibull AFT applied to lung data. Second figure : Cox's Proportional Hazards model applied to same lung data.

3.2. **The Proportional Hazards Assumption.**

A key assumption of Cox's regression model is the *proportional hazards assumption*. This assumption states that the ratios of any two vector covariate groups are independent of time [9, page 245]. This can be seen by comparing the hazard functions of two individuals a and b with covariate vectors $\boldsymbol{x^a} = (x_1^a, ..., x_p^a)$ and $\boldsymbol{x^b} = (x_1^b, ..., x_p^b)$ respectively:

[5]Although the second graph in figure 2 appears to be similar to a Kaplan-Meier estimated survival curve, it is not estimated by these means. `R` plots Cox PH survival curves through predictor survival functions for a Cox PH model, see [9, section 8.8] for more details and (10) for an idea as to how this works.

$$\frac{\lambda(t|\boldsymbol{X^a})}{\lambda(t|\boldsymbol{X^b})} = \frac{\lambda_0(t)\exp\left(\sum_{j=1}^{p}\beta_j x_j^a\right)}{\lambda_0(t)\exp\left(\sum_{j=1}^{p}\beta_j x_j^b\right)} = \exp\left(\sum_{j=1}^{p}\beta_j(x_j^a - x_j^b)\right).$$

This expression is seen to be constant, which implies that the hazard functions at any point in time between any two individuals in a time-to-event data set are proportional. This assumption is quite strong so would ideally be tested, for which there are a few methods of doing (see section 4). The covariates in this model are said to be *time-independent*, such that they do not change over time. Time-dependent variables would violate the proportional hazards assumption, however a Cox regression can still be applied to time-dependent variables under an *extended Cox model*. Time-dependent covariates are not discussed in this project, however the reader can study chapter 6 of Kleinbaum and Klein Survival Analysis text[8] or chapter 9 of Klein and Moeschberger's text[9] for further details.

3.3. **Estimation of Regression Coefficients.**

The coefficients $\boldsymbol{\beta}$ are the parameters of interest when applying Cox's regression model to data. However, estimation of these coefficients proves difficult due to the infinite-dimensional nuisance baseline hazard $\lambda_0(t)$. Cox found a way around this problem though, resulting in the *partial likelihood* method for obtaining the parameter estimates $\hat{\boldsymbol{\beta}}$. We first investigate why the baseline hazards prevent us from estimating $\boldsymbol{\beta}$ through a conventional full likelihood.

Hosmer, Lemeshow and May attempted to derive $\boldsymbol{\beta}$ through the full likelihood function for Cox's model by means of maximum likelihood as follows [12, pages 72-74]:
Assume we are interested in finding the maximum likelihood estimate of coefficient $\hat{\beta}$ of only single covariate x. Let the triplet (t_i, x_i, c_i) denote the lifetime, the covariate (fixed throughout the study) and censoring status, respectively, of the i^{th} individual of a population of n independent observations, where

$$c_i = \begin{cases} 1, & \text{if } i^{th} \text{ invidual has an uncensored survival time;} \\ 0, & \text{if } i^{th} \text{ invidual has a right-censored survival time.} \end{cases}$$

The likelihood function is constructed through formulating an expression for determining the contribution of all individuals effect on this likelihood. For censored observations, all we know is that these observations has survival time atleast t_i. Therefore, their contribution to the likelihood function is given by $S(t, x, \beta)$. For uncensored observation, we know the exact survival time. Therefore, their contribution to the likelihood function is given by $f(t, x, \beta)$[6] . The likelihood function is therefore given by the product of each individuals contribution:

$$L(\beta) = \prod_{i}^{n} \left\{ [f(t_i, x_i, \beta)]^{c_i} [S(t_i, x_i, \beta)]^{1-c_i} \right\}. \tag{4}$$

The parameter estimate is then found by maximising function (4) which is done more easily by maximising the log-likelihood of (4), yielding equivalent estimates of β. The

[6] f, as seen in (2), is the function with $f(t)\,dt$ being the probability that time t the event of interest happens.

log-likelihood function $l(\beta)$ is given by:

$$\begin{aligned} l(\beta) &= \ln(L(\beta)) \\ &= \sum_{i=1}^{n} \ln\left\{ \left[f(t_i, x_i, \boldsymbol{\beta})\right]^{c_i} \left[S(t_i, x_i, \boldsymbol{\beta})\right]^{1-c_i} \right\} \\ &= \sum_{i=1}^{n} \left\{ c_i \ln\left[f(t_i, x_i, \boldsymbol{\beta})\right] + (1-c_i) \ln\left[S(t_i, x_i, \boldsymbol{\beta})\right] \right\}. \end{aligned} \tag{5}$$

By substituting the expression $f(t_i, x_i, \beta) = S(t_i, x_i, \beta)\lambda(t_i, x_i, \beta)$ (yielded from (2)) into (5) and using the expression $S(t) = S_0(t)^{\exp(\sum_{j=1}^{p} \beta_j x_j)}$ (see (10) for proof) we obtain:

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \left\{ c_i \ln\left[S(t_i, x_i, \beta)\lambda(t_i, x_i, \beta)\right] + (1-c_i) \ln\left[S_0(t_i)^{\exp(\beta x)}\right] \right\} \\ &= \sum_{i=1}^{n} \left\{ c_i \ln\left[S_0(t_i)^{\exp(\beta x)}\right] + c_i \ln\left[\lambda_0(t_i)\exp(\beta x)\right] + (1-c_i) \ln\left[S_0(t_i)^{\exp(\beta x)}\right] \right\} \\ &= \sum_{i=1}^{n} \left\{ c_i \ln\left[\lambda_0(t_i)\right] + c_i x_i \beta + \exp(\beta x_i) \ln\left[S_0(t_i)\right] \right\}. \end{aligned} \tag{6}$$

Kalbfleisch and Prentice demonstrated[17] why the unspecified baseline hazards $S_0(t_i)$ and $\lambda_0(t_i)$ halts the process of finding the estimate for β by these means in their text *The Statistical Analysis of Failure Time Data*.
Cox first proposed a technique to tackle this problem and estimate $\boldsymbol{\beta}$ by means of a method referred to as *conditional likelihood*, in the paper Regression Models and Life-Tables[4]. This method involved removing the baseline hazard parameter from the equation estimating $\boldsymbol{\beta}$ by assuming it is arbitrary. However, this version of Cox's method soon became under scrutiny as others came to the conclusion that what Cox suggested was not a conditional likelihood. Cox gave a more detailed justification when in 1975 he published a revised method, *partial likelihood*[11]. The details of the theory is as follows:

Suppose in a population of n individuals with survival times $T_1 < ... < T_n$ there exists k observed (uncensored) event times $t_1 < ... < t_k$ with $k \leq n$. Let:

- x_j^a be the j^{th} covariate for the a^{th} individual;
- $R(t_i)$ the set of all individuals alive just before event time t_i, whom are at risk of experiencing the event of interest.

The partial-likelihood function $L(\beta)$, in words, is the product of individual likelihoods $L_1, .., L_k$ for each event time $t_1, ..., t_k$. Each of these individual likelihoods are the ratios of the instantaneous rate of failure for the individual with this specific event time, to the cumulative rate of failure for all individuals at risk of experiencing this event time (which includes the individual who experiences the event of interest at this event time).

Assuming that censoring is non-informative, such that what determines censoring is unrelated to event of the interest[8, page 442 for more information] and that there are no tied survival times of individuals, the partial likelihood is given by[7] [8, page 99]:

[7]Whilst we have not included censored observations in our partial likelihood, some authors (such as Hosmer, Lemeshow and May[12]) express the partial likelihood function including censored observations, which is given equally as:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\sum_{j=1}^{p} \beta_j x_j^i)}{\sum_{\alpha \in R(t_i)} \exp(\sum_{j=1}^{p} \beta_j x_j^\alpha)} \right]^{c_i}$$

$$
\begin{aligned}
L(\beta) &= L_1 \times L_2 \times ... \times L_k \\
&= \frac{\text{Hazard function of individual with event time 1}}{\text{Summation of all hazard functions for individuals at risk at event time 1}} \times ... \\
&... \times \frac{\text{Hazard function of individual with event time k}}{\text{Summation of all hazard function for individuals at risk at event time k}} \\
&= \frac{\lambda_0(t)\exp\left(\sum_{j=1}^{p}\beta_j x_j^1\right)}{\sum_{\alpha\in R(t_1)}\lambda_0(t)\exp\left(\sum_{j=1}^{p}\beta_j x_j^\alpha\right)} \times ... \times \frac{\lambda_0(t)\exp\left(\sum_{j=1}^{p}\beta_j x_j^k\right)}{\sum_{\alpha\in R(t_k)}\lambda_0(t)\exp\left(\sum_{j=1}^{p}\beta_j x_j^\alpha\right)} \\
&= \prod_{i=1}^{k}\frac{\exp(\sum_{j=1}^{p}\beta_j x_j^i)}{\sum_{\alpha\in R(t_i)}\exp(\sum_{j=1}^{p}\beta_j x_j^\alpha)}. \qquad (7)
\end{aligned}
$$

From (7) we can see that the baseline hazards have cancelled out from the equation. We have found a solution to the problem found by Kalbfleisch and Prentice concerning estimating the likelihood with baseline hazards!
The estimates for $\boldsymbol{\beta}$ are then computed by maximising the natural log of $L(\boldsymbol{\beta})$. Let $l(\boldsymbol{\beta}) = \ln(L(\boldsymbol{\beta}))$, we can rewrite (7) as:

$$
\begin{aligned}
l(\boldsymbol{\beta}) &= \ln\left[\frac{\exp(\sum_{j=1}^{p}\beta_j x_j^1)\times ... \times \exp(\sum_{j=1}^{p}\beta_j x_j^k)}{\sum_{\alpha\in R(t_1)}\exp(\sum_{j=1}^{p}\beta_j x_j^\alpha)\times ... \times \sum_{\alpha\in R(t_k)}\exp(\sum_{j=1}^{p}\beta_j x_j^\alpha)}\right] \\
&= \ln\left[\exp(\sum_{j=1}^{p}\beta_j x_j^1)\times ... \times \exp(\sum_{j=1}^{p}\beta_j x_j^k)\right] - \ln\left[\sum_{\alpha\in R(t_1)}\exp(\sum_{j=1}^{p}\beta_j x_j^\alpha)\times ... \times \sum_{\alpha\in R(t_k)}\exp(\sum_{j=1}^{p}\beta_j x_j^\alpha)\right] \\
&= \sum_{i=1}^{k}\sum_{j=1}^{p}\beta_j x_j^i - \sum_{i=1}^{k}\ln\left[\sum_{\alpha\in R(t_i)}\exp\left(\sum_{j=1}^{p}\beta_j x_j^\alpha\right)\right]. \qquad (8)
\end{aligned}
$$

Taking the first derivative of (8) with respect to β_a, $\forall a \in \{1,...,p\}$, we find that:

$$
\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_a} = \sum_{i=1}^{k} x_a^i - \sum_{i=1}^{k}\left[\frac{1}{\sum_{\alpha\in R(t_i)}\exp\left(\sum_{j=1}^{p}\beta_j x_j^\alpha\right)}\cdot\sum_{\alpha\in R(t_i)} x_a^\alpha \exp\left(\sum_{j=1}^{p}\beta_j x_j^\alpha\right)\right]. \qquad (9)
$$

The estimated coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)$ are then calculated by solving the p-system of equations given by setting (9) equal to 0: $\frac{\partial l(\beta)}{\partial \boldsymbol{\beta}_a} = 0$, $\forall a \in \{1,...,p\}$.
Cox put forward the notion that his proposed partial-likelihood method of obtaining the parameter estimates has similar properties to the maximum likelihood method described in (6). This was demonstrated to be true by Andersen et al. in their text *Statistical Models Based on Counting Processes*[13] and Fleming and Harrington in their text *Counting Processes and Survival Analysis*[14].
Due to the computationally heavy nature of this task, this is usually achieved by iterative methods (such as the Newton-Raphson technique) using software environments such as R, SPSS or MATLAB.

3.4. **Tied survival times.**
In the previous subsection, the partial-likelihood method of obtaining parameter estimates $\hat{\boldsymbol{\beta}}$ had a crucial assumption that the k event times $t_1 < ... < t_k$ were distinct.
If we consider two individuals A and B who share an equal event time and apply Cox's partial-likelihood procedure previously described in (7), we are left with an interesting question: do we consider individual A as the individual experiencing the event of interest and individual B in the risk group, or vice-versa?
In reality, the assumption of k distinct event times is not likely satisfied, as survival times are usually grouped on a discrete time frame (such as days or months). In the circumstance in which there exists tied survival times between individuals in a population, a few methods have been proposed.

For the entirety of the remaining subsection, let

- $t_1 < ... < t_k$ denote uncensored (observed) survival times;
- $i_1, ..., i_{m_i}$ denote the individuals sharing the event time t_i;
- Q_i be the set of permutations of individuals, $i_1, ..., i_{m_i}$, with shared event time t_i;
- $P = (p_1, .., p_{m_i})$ be one element in Q_i;
- $R(t_i)$ be the set of individuals at risk of experiencing the event at time t_i;
- $R(t_i, P, k) = R(t_i) - (p_1, ..., p_{r-1})$;
- $\boldsymbol{S}^i = (S_1^i, ..., S_p^i) = (\sum_{i=1}^{m_i} x_1^i, ..., \sum_{i=1}^{m_i} x_p^i)$.

Breslow suggested adjusting the partial-likelihood defined by Cox (7) when dealing with tied-times by taking each individual likelihood $L_1, .., L_k$ to be the ratio of the instantaneous rate of failure of an individual with covariates equal to the sum of all relative covariates of individuals with this event time, to the cumulative hazard of all individuals at risk at this event time to the power of the number of tied-times at this event time[15]:

$$L_B(\beta) = \prod_{i=1}^{k} \frac{\exp\left(\sum_{j=1}^{p} \beta_j S_j^i\right)}{\left[\sum_{\alpha \in R(t_i)} \exp\left(\sum_{j=1}^{p} \beta_j x_j^\alpha\right)\right]^{m_i}}.$$

Breslow's method works well if the ratio, for each event time, of the number of individuals with this event time to the number of individuals at risk at this event time is small.

In such circumstances for which these ratios are not small, Efron suggested another alternative likelihood function[16]:

$$L_E(\beta) = \prod_{i=1}^{m} \frac{\exp\left(\sum_{j=1}^{p} \beta_j S_j^i\right)}{\prod_{k=1}^{m_i} \left(\sum_{\alpha \in R(t_i)} \exp\left(\sum_{j=1}^{p} \beta_j x_j^\alpha\right) - \frac{k-1}{m_i} \sum_{\alpha \in D(t_i)} \exp\left(\sum_{j=1}^{p} \beta_j x_j^\alpha\right)\right)}$$

Efron's method is the default method for dealing with tied-times used by the survival package in `R`.

Prentice and Kalbfleisch described an exact method[17]:

$$L_{PK}(\beta) = \prod_{i=1}^{k} \left[\frac{\exp\left(\sum_{j=1}^{p} \beta_j S_j^i\right)}{\prod_{r=1}^{m_i} \left\{\sum_{\alpha \in R(t_i, P, k)} \exp\left(\sum_{j=1}^{p} \beta_j x_j^\alpha\right)\right\}} \right]$$

Prentice and Kalbfleisch's method, through considering the number of permutations that can arise from large numbers of m_i, can become computationally heavy.

Hertz-Picciotto and Rockhill analysed and compared these 3 different methods for dealing with tied-times in their article *Validity and Efficiency of Approximation Methods for Tied Survival Times in Cox Regression*[18]. Their results concluded that Efron's method was the most optimal method for determining the true value of beta.

4. Testing the Proportional Hazards assumption

There exists two primary graphical diagnostic methods for determining whether the proportional hazards assumption has been satisfied: comparison of observed and predicted survival curves; and comparison of log-log curves. Evaluation of Schoenfield residuals poses a more practical approach to diagnosis of the proportional hazards assumption. In this section, we explore these methods.

4.1. **Log-Log curves.**

In determining how log-log curves help us in diagnosing the proportional hazards assumption, we first consider how the survival function is related to the cumulative hazard function and then find an expression for the survival function in Cox's model.

Let $\Lambda(t) = \int_0^t \lambda(u)\,du$ be the cumulative hazard function. It can be shown that survival functions are related to cumulative hazard functions in the following way:
It has been shown in (2) that

$$\lambda(t) = \frac{-S'(t)}{S(t)} = -\frac{d\log((S(t))}{dt}.$$

Hence,

$$\begin{aligned}
\Lambda(t) &= \int_0^t \lambda(u)\,du \\
&= -\int_0^t \frac{d\log(S(u))}{du}\,du \qquad \textit{(Substituting alternative definition of hazard function)} \\
&= -\log(S(t)). \\
&\iff \exp(-\Lambda(t)) = S(t).
\end{aligned}$$

Thus, we can express the survival function in the context of Cox's model as:

$$\begin{aligned}
S(t) &= \exp(-\Lambda(t)) \\
&= \exp\left(-\int_0^t \lambda(u)\,du\right) \qquad \textit{(Definition of cumulative hazard)} \\
&= \exp\left(-\int_0^t \lambda_0(u)\exp(\sum_{j=1}^{p}\beta_j x_j)\,du\right) \qquad \textit{(Using formula for Cox's hazard function)} \\
&= \exp\left(-\int_0^t \lambda_0(u)\,du\right)^{\exp(\sum_{j=1}^{p}\beta_j x_j)} \\
&= S_0(t)^{\exp(\sum_{j=1}^{p}\beta_j x_j)}. \qquad (10)
\end{aligned}$$

A graphical interpretation of the proportional-hazards assumption being satisfied are log-log curves of these survival functions. When taking the logarithm of the negative

logarithm of the survival function defined in (10), we find [8, page 139]:

$$\begin{aligned}\log(-\log(S(t)) &= \log\left[-\log(S_0(t)^{\exp(\sum_{j=1}^{p}\beta_j x_j)})\right] \\ &= \log\left[-\exp(\sum_{j=1}^{p}\beta_j x_j)\log(S_0(t))\right] \\ &= \log\left[-\log(S_0(t))\right] + \sum_{j=1}^{p}\beta_j x_j.\end{aligned}$$

Thus for two individuals a and b with covariate vectors $\boldsymbol{x^a} = (x_1^a, ..., x_p^a)$ and $\boldsymbol{x^b} = (x_1^b, ..., x_p^b)$ respectively, we can see that the distance between each $\log(-\log(S(t))$ curves for each individual is constant over time:

$$\begin{aligned}\log\left[-\log(S_0(t))\right] + \sum_{j=1}^{p}\beta_j x_j^a - \log\left[-\log(S_0(t))\right] - \sum_{j=1}^{p}\beta_j x_j^b &= \sum_{j=1}^{p}\beta_j x_j^a - \sum_{j=1}^{p}\beta_j x_j^b \\ &= \sum_{j=1}^{p}\beta_j(x_j^a - x_j^b).\end{aligned}$$

Figure 3 displays the log-log survival curves for males and females of the lung cancer

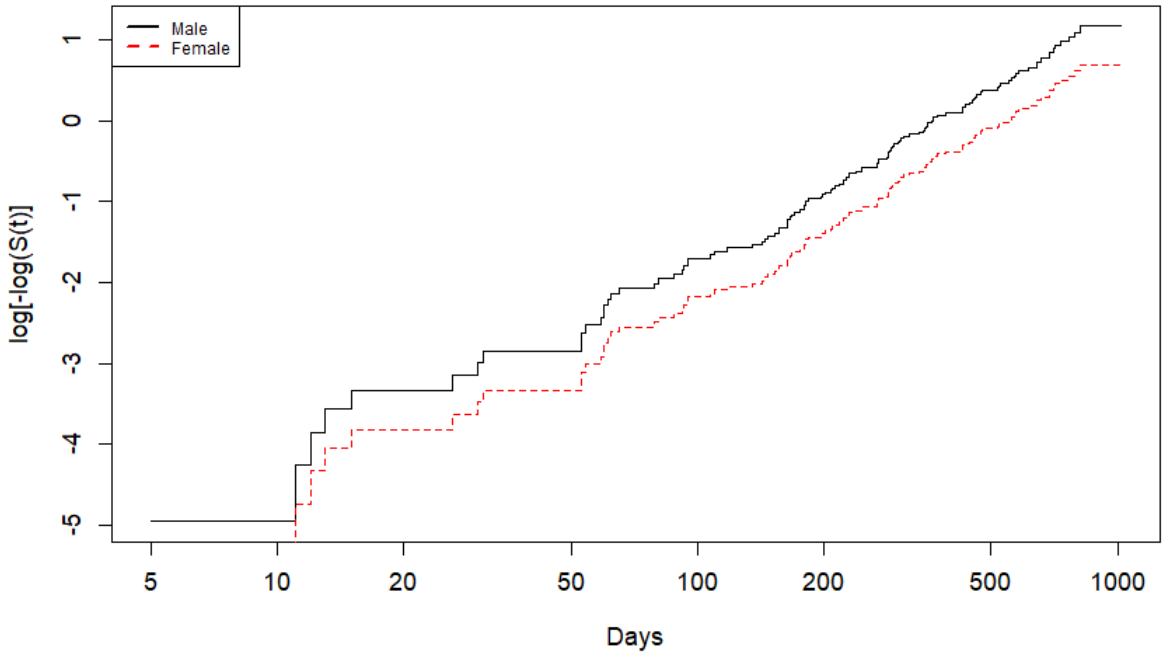


FIGURE 3. $\log(-\log(S(t))$ plots of survival curves for male and females.

data. The distance between the two curves seems to be constant on the whole, with a difference of $\sum_{j=1}^{p}\beta_j(x_j^a - x_j^b) \approx \sum_{j=1}^{1} -0.447(1-2) = 0.447$. We can therefore conclude that the covariate sex does not violate the proportional hazards assumption by this method.

4.2. **Comparison of Kaplan-Meier and Cox PH regression estimated survival curves.**

An alternative graphical method for testing the proportional hazards assumption involves

comparison of estimated survival curves. We will compare the estimated survival curves obtained from fitting a Cox PH regression to the expected survival curves obtained from using the Kaplan-Meier estimator.
When dealing with multiple covariates, this is usually done by dealing with one covariate at a time. The covariate being investigated has survival curves plotted for all values which the covariate can take. If the covariate is continuous, it is discretized and then survival curves are plotted for each category of the discretized continuous covariate. The covariates are presumed to not violate the proportional hazards assumption if they are similar. Klein and Kleinbaum put forward the notion that plots which are close comply with the PH assumption, and discrepancies between plots violate the PH assumption [8, page 147].
Figure 4 below shows this method of testing the proportional hazards assumption for the covariate sex, in the lung cancer data we have been following. For the most part, we observe that the plots for each estimated survival curves for both males and females are close and therefore we do not reject the proportional hazards assumption.

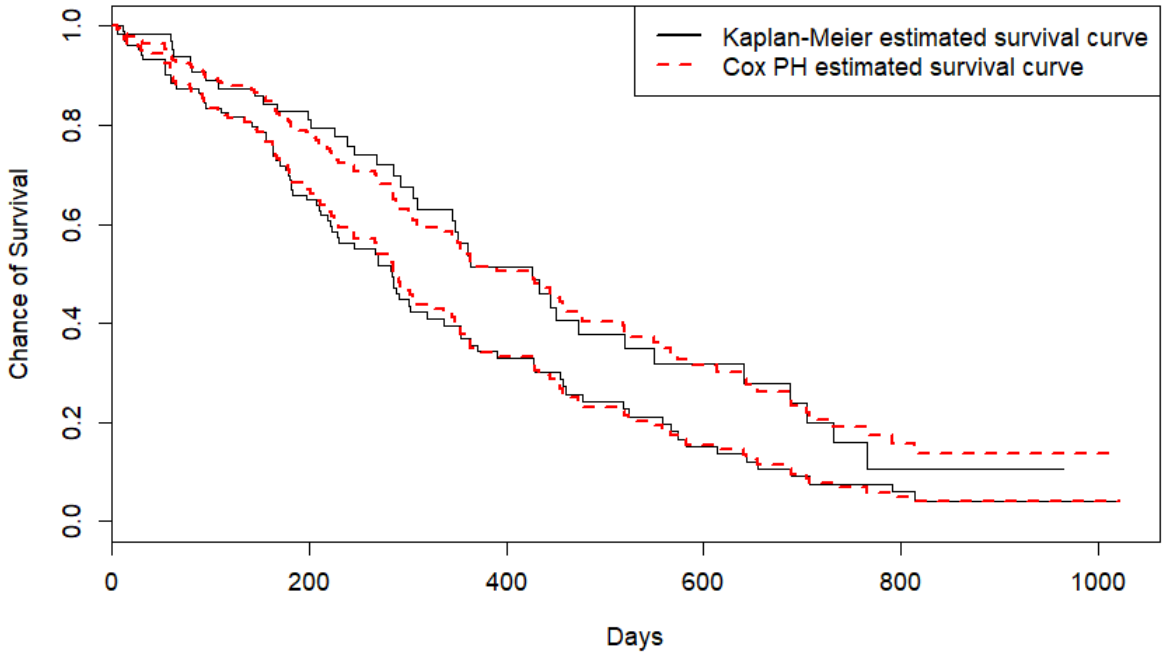


FIGURE 4. Comparison of estimated survival curves between Kaplan-Meier estimates and Cox regression for covariate sex in lung data. Plots show survival curves for males and females.

4.3. **Schoenfield Residuals.**
A more practical approach in determining whether a proposed Cox's model satisfies the proportional hazards assumption is using Schoenfield residuals. The use of Schoenfield residuals in determining the validation of the PH assumption were first demonstrated by Harrell and Lee in their article Evaluating the Yield of Medical Tests[19]. Schoenfield defined the partial residuals as follows[20]:

As usual, let the i^{th} individual of a time-to-event data set of size n have covariate vector $\boldsymbol{x^i} = (x_1^i, ..., x_p^i)$ and experience an event time t_i. Let k of these individuals have distinct observed (uncensored) event times $t_1 < ... < t_k$ $(k \leq n)$ and the individuals at risk at event time t_i be denoted $R(t_i)$.
In the derivation of the first partial derivative of Cox's partial likelihood with respects to covariate a shown in (9), we found that

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_a} = \sum_{i=1}^{k} x_a^i - \sum_{i=1}^{k} \left[\frac{1}{\sum_{\alpha \in R(t_i)} \exp\left(\sum_{j=1}^{p} \beta_j x_j^\alpha\right)} \cdot \sum_{\alpha \in R(t_i)} x_a^\alpha \exp\left(\sum_{j=1}^{p} \beta_j x_j^\alpha\right) \right].$$

This can be rewritten as

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_a} = \sum_{i=1}^{k} \left[x_a^i - \frac{\sum_{\alpha \in R(t_i)} x_a^\alpha \exp\left(\sum_{j=1}^{p} \beta_j x_j^\alpha\right)}{\sum_{\alpha \in R(t_i)} \exp\left(\sum_{j=1}^{p} \beta_j x_j^\alpha\right)} \right]. \tag{11}$$

The fraction on the right-hand side of equation (11) can be considered as a weighted average of covariate x_a conditional on the risk set $R(t_i)$, with weights being the hazard rates of all individuals in this risk set. We therefore let

$$\frac{\sum_{\alpha \in R(t_i)} x_a^\alpha \exp\left(\sum_{j=1}^{p} \beta_j x_j^\alpha\right)}{\sum_{\alpha \in R(t_i)} \exp\left(\sum_{j=1}^{p} \beta_j x_j^\alpha\right)} = \mathbb{E}[x_\alpha^i | R(t_i)].$$

Equation (11) becomes

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_a} = \sum_{i=1}^{k} \left[x_a^i - \mathbb{E}[x_a^i | R(t_i)] \right]. \tag{12}$$

Define the partial Schoenfield residual at event time t_i as

$$\hat{\boldsymbol{r}}_{\boldsymbol{i}} = (\hat{r}_1^i, ..., \hat{r}_p^i)$$

with,

$$\hat{r}_a^i = x_a^i - \hat{\mathbb{E}}[x_a^i | R(t_i)].$$

where $\hat{\mathbb{E}}[x_a^i | R(t_i)]$ is obtained from substituting the computed coefficient estimates $\hat{\boldsymbol{\beta}}$ into $\mathbb{E}[x_\alpha^i | R(t_i)]$ for $\boldsymbol{\beta}$.[8]
The Schoenfield residual $\hat{r}_\alpha^i$ for the i^{th} individual and a^{th} covariate is therefore the difference between the observed value of x_α^i and the expectation of x_α^i given $R(t_i)$.

It can be seen that in a population of n individuals with k event times and p covariates, there exists kp Schoenfield residuals.
Harrell and Lee proposed that the proportional hazards assumption would not be rejected for covariate $i \in \{1, ..., p\}$ if, when plotted against time, $\hat{\boldsymbol{r}}_{\boldsymbol{i}}$ displayed no correlation [19].
Hypotheses tests which test the null hypothesis $H_0 : \rho = 0$ against the alternative hypothesis $H_A : \rho \neq 0$, with ρ representing the correlation coefficient of Schoenfield residuals for each covariate are formal methods of determining whether correlation exists.
Grambsch and Therneau proposed that scaled Schoenfield residuals would be of more use [22], which are used by software environments such as R and SPSS when computing the residuals.
The plotted scaled Schoenfield residuals against time for covariates age and sex are shown in Figure 5, a smooth red line that does not change over time indicates that the proportion hazards assumption is not violated for covariate sex or age. Table 5 verifies this,

[8]Since $\hat{\boldsymbol{\beta}}$ is indeed the solution to (12), we would expect the sum of the Schoenfield residuals $\hat{r}_a^1, \cdots, \hat{r}_a^k$ to be 0.

TABLE 5. Sample correlation coefficient and p-values for Schoenfield residuals of covariates sex and age.

Covariate	ρ	p-value
Sex	0.0841	0.366
Age	-0.0633	0.483
Global	NA	0.486

with the reported correlation coefficient values ρ being close to 0 and p-values being large, we do not reject $H_0 : \rho = 0$, the scaled Schoenfield residuals do not vary with time and therefore the proportional hazards assumption is not rejected.

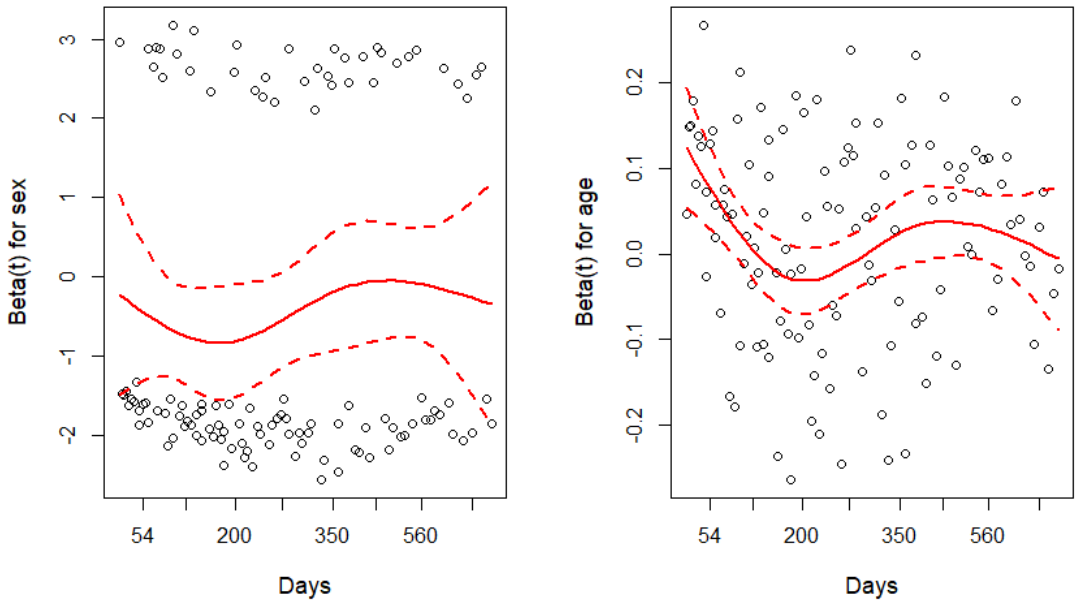


FIGURE 5. Scaled Schoenfield residuals plotted against time for covariates sex (left) and age (right).

5. Tests for Significance of Predictor Variables

This section outlines three methods for testing the goodness of fit of a Cox PH regression model: Wald's test, the likelihood ratio test and the log rank test. All three of these tests are asymptotically equivalent. This means that, given suitably significant covariates with a large enough sample, all tests would converge to the same result. These tests use a mixture of the *efficient score vector*, *Fisher's information* and the *observed information matrix*. Cox and Hinkley give detailed notes on the theory of these functions in their text *Theoretical Statistics* [21, chapter 9]. We first define these functions and derive these functions for Cox's partial likelihood.

5.1. The Efficient Score Vector, Fisher's Information and the Observed Information Matrix for Cox's Partial Likelihood Function.

Let $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ with likelihood function $L(\boldsymbol{\beta})$ and log-likelihood function $l(\boldsymbol{\beta}) =$

$\ln L(\boldsymbol{\beta})$. The *efficient score vector* is denoted $\boldsymbol{S(\beta)} = (S_1(\boldsymbol{\beta}), ..., S_p(\boldsymbol{\beta}))$, where $S_a(\boldsymbol{\beta})$ is given by [9, page 449]

$$S_a(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_a}, \quad \forall a \in \{1, ..., p\}.$$

Fisher's information matrix $\boldsymbol{i_\beta}$ is given by the elements [9, page 450]

$$\begin{aligned} i_{a,b}(\boldsymbol{\beta}) &= \mathbb{E}\left[-\frac{\partial \boldsymbol{S(\beta)}}{\partial \boldsymbol{\beta}}\right] \\ &= -\mathbb{E}\left[\frac{\partial^2}{\partial\beta_a \partial\beta_b} l(\boldsymbol{\beta})\right], \quad \forall a, b \in \{1, ..., p\}. \end{aligned}$$

A consistent estimator of $\boldsymbol{i_\beta}$ is the *observed information* $\boldsymbol{I_\beta}$, which $(a,b)^{th}$ element is given by [9, page 450]

$$I_{a,b}(\boldsymbol{\beta}) = -\frac{\partial l(\boldsymbol{\beta})}{\partial\beta_a\beta_b}. \tag{13}$$

We now find the equivalent of these functions for Cox's partial likelihood. The efficient score vector for Cox's PH model[9] was derived at (9), which is rewritten as

$$S_a(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_a} = \sum_{i=1}^{k}\left[x_a^i - \frac{\sum_{\alpha\in R(t_i)} x_a^\alpha \exp\left(\sum_{j=1}^p \beta_j x_j^\alpha\right)}{\sum_{\alpha\in R(t_i)} \exp\left(\sum_{j=1}^p \beta_j x_j^\alpha\right)}\right]. \tag{14}$$

The observed information matrix of the likelihood function of Cox's regression is derived as follows:
We need to find the second derivative of (9) with respects to β_b. We split up (9) into two functions $Q(\beta_b)$ and $F(\beta_b)$ as follows:

$$\frac{\partial l(\beta)}{\partial \beta_a} = \underbrace{\sum_{i=1}^{k} x_a^i}_{Q(\beta_b)} - \underbrace{\sum_{i=1}^{k}\left[\frac{1}{\sum_{\alpha\in R(t_i)} \exp\left(\sum_{j=1}^p \beta_j x_j^\alpha\right)} \cdot \sum_{\alpha\in R(t_i)} x_a^\alpha \exp\left(\sum_{j=1}^{p} \beta_j x_j^\alpha\right)\right]}_{F(\beta_b)}.$$

$Q(\beta_b)$ vanishes when taking the derivative with respects to β_b. When computing the derivative of $F(\beta_b)$ we shall use the quotient rule, and set:

$$F(\beta_b) = \frac{g(\beta_b)}{h(\beta_b)} = \frac{\sum_{i=1}^k \left[\sum_{\alpha\in R(t_i)} x_a^\alpha \exp\left(\sum_{j=1}^p \beta_j^\alpha\right)\right]}{\sum_{i=1}^k \left[\sum_{\alpha\in R(t_i)} \exp\left(\sum_{j=1}^p \beta_j^\alpha\right)\right]}.$$

obtaining,

$$\begin{aligned} \frac{\partial F(\beta_b)}{\partial \beta_b} &= \frac{g'(\beta_b)h(\beta_b) - g(\beta_b)h'(\beta_b)}{(h(\beta_b))^2} \\ &= \frac{g'(\beta_b)h(\beta_b)}{(h(\beta_b))^2} - \frac{g(\beta_b)h'(\beta_b)}{(h(\beta_b))^2}. \end{aligned}$$

[9] The efficient score vector is seen to be a $1 \times p$ vector.

We find that

$$g'(\beta_b) = \sum_{i=1}^{k}\left[\sum_{\alpha\in R(t_i)} x_a^\alpha x_b^\alpha \exp(\sum_{j=1}^{p}\beta_j x_j^\alpha)\right],$$
$$h'(\beta_b) = \sum_{i=1}^{k}\left[\sum_{\alpha\in R(t_i)} x_b^\alpha \exp(\sum_{j=1}^{p}\beta_j x_j^\alpha)\right].$$

Plugging these expressions into (13), we find the observed information matrix[10]:

(15)

$$-\frac{\partial^2 l(\beta)}{\partial\beta_a\beta_b} = \sum_{i=1}^{k}\left[\frac{\sum_{\alpha\in R(t_i)} x_a^\alpha x_b^\alpha \exp(\sum_{j=1}^{p}\beta_j x_j^\alpha)}{\sum_{\alpha\in R(t_i)}\exp(\sum_{j=1}^{p}\beta_j x_j^\alpha)}\right]$$
$$-\sum_{i=1}^{k}\left[\frac{\sum_{\alpha\in R(t_i)} x_a^\alpha \exp(\sum_{j=1}^{p}\beta_j x_j^\alpha)\sum_{\alpha\in R(t_i)} x_b^\alpha \exp(\sum_{j=1}^{p}\beta_j x_j^\alpha)}{(\sum_{\alpha\in R(t_i)}\exp(\sum_{j=1}^{p}\beta_j x_j^\alpha))^2}\right].$$
$$= \boldsymbol{I}_{a,b}(\boldsymbol{\beta})$$

5.2. **Wald's Test.**
Wald's test is the main hypothesis test in determining the significance of regression parameters. In the context of Cox's PH regression model, let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)$ denote the p maximum likelihood estimates coefficients for the p covariates being tested and $\boldsymbol{I}_{a,b}(\boldsymbol{\beta})$ denote the observed information matrix in (15). Under circumstances with large samples, $\hat{\boldsymbol{\beta}} \sim \boldsymbol{N_p}(\boldsymbol{\beta}, I^{-1}(\hat{\boldsymbol{\beta}}))$[9, page 254].

With hypotheses

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0} \text{ Vs. } H_A : \boldsymbol{\beta} \neq \boldsymbol{\beta_0}$$

and test statistic X_W^2,

$$X_W^2 = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})' I(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})$$

where,

$$X_W^2 \overset{H_0}{\sim} \chi_p^2.$$

We reject H_0 at significance level α if we observe

$$X_W^2 > \chi_p^2(1-\alpha).$$

where $X_p^2(1-\alpha)$ denotes the $(1-\alpha)$ quantile of the Chi-squared distribution with p degrees of freedom.

[10]The observed information matrix obtained is seen to be a $p \times p$ matrix.

5.3. **Likelihood Ratio Test.**
An alternative to Wald's test is the likelihood ratio test. Under circumstances with large samples, it tests the hypotheses [9, page 254].

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0} \text{ Vs. } H_A : \boldsymbol{\beta} \neq \boldsymbol{\beta_0}$$

and test statistic X^2_{LR},

$$X^2_{LR} = 2[l(\hat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta_0})]$$

where,

$$X^2_{LR} \overset{H_0}{\sim} \chi^2_p.$$

We reject H_0 at significance level α if we observe

$$X^2_{LR} > \chi^2_p(1-\alpha).$$

where $X^2_p(1-\alpha)$ denotes the $(1-\alpha)$ quantile of the Chi-squared distribution with p degrees of freedom.
In the context of Cox's model,

- $l(\hat{\boldsymbol{\beta}})$ is the maximised log-partial-likelihood of the model involving estimated coefficients.
- $l(\boldsymbol{\beta_0})$ is the maximised log-partial-likelihood of the model involving coefficients we wish to test against.

5.4. **Score Test.**
The score test is based on the efficient score vector $\boldsymbol{S(\beta)}$. For a large amount of n observations, $\boldsymbol{S(\beta)} \sim \boldsymbol{N_p(0, I(\beta))}$[9, page 255]. It tests the hypotheses

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0} \text{ Vs. } H_A : \boldsymbol{\beta} \neq \boldsymbol{\beta_0}\,.$$

It has test statistic X^2_S,

$$\chi^2_S = \boldsymbol{S(\beta_0) I^{-1}(\beta_0) S'(\beta_0)}$$

where,

$$X^2_S \overset{H_0}{\sim} \chi^2_p.$$

We reject H_0 at significance level α if we observe

$$X^2_S > \chi^2_p(1-\alpha).$$

where $X^2_p(1-\alpha)$ denotes the $(1-\alpha)$ quantile of the Chi-squared distribution with p degrees of freedom.
In the context of Cox's model, we set $\boldsymbol{S(\beta)}$ to be the efficient score vector derived in (14).

Usually, $\boldsymbol{\beta_0}$ is set to $\mathbf{0}$. Under the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$, each of these 3 tests investigates the significance of the coefficients.

Table 6 displays the test statistics and corresponding p-values for each test described above for the Cox PH model involving the covariates sex and age. For an $\alpha = 0.05$ significance level, we can reject $H_0 : \boldsymbol{\beta} = \mathbf{0}$ for all three tests and conclude that, for the

TABLE 6. Likelihood ratio, Wald's test and Score test results for Cox model fitting covariates sex and age.

Test	Test statistic value	p-value
Likelihood ratio	8.88	0.01179
Wald's	8.47	0.01449
Score (log-rank)	8.6	0.0136

lung data with covariates sex and age in the model, at least one of the coefficients is not 0.

6. Stratification of covariates which violate the PH assumption

After discussing methods of investigating whether covariates satisfy the crucial proportional hazard assumption in section 4, we now explore approaches to dealing with such covariates. In circumstances which covariates violate the PH assumption, such covariates cannot be included in the proposed Cox PH model discussed up to this point. In this section an extension of the original Cox PH model, called the *Stratified Cox* (SC) PH model, is investigated. The SC PH model outlines a method for dealing with these covariates which do not comply with the proportional hazards assumption. Kleinbaum and Klein present the SC PH procedure in a coherent manner, which this next section is based upon[8, pages 176-188].

6.1. **The Stratified Cox Proportional Hazards Model.**

Suppose we are applying a Cox PH model to a group of p covariates, of which n satisfy the PH assumption and k do not. Denote these covariates as follows:

$$\underbrace{x_1, x_2, ..., x_n,}_{n \text{ PH satisfying covariates}} \qquad \underbrace{y_1, y_2, ..., y_k.}_{k \text{ PH violating covariates}}$$

Define a new variable Y with G groups. These groups are every combination of categories of all the stratified k PH violating covariates.
As an example, suppose in a fitted Cox PH model we find that there are $k = 2$ violating PH covariates y_1 and y_2. The first covariate y_1 is a dichotomous variable, with only two categories. The second covariate y_2 is a continuous covariate which is discretized into 3 categories. In this case, the new variable Y would have $G = 6$ categories based on the 2 stratified PH violating covariates with 3 and 2 categories, one for each combination.

The hazard function for a SC PH model is given by:

$$\lambda^g(t) = \lambda_0^g(t) \exp(\sum_{j=1}^{n} \beta_j x_j) \quad \text{for each group g=1,...,G.} \tag{16}$$

Note that:

- There are different hazard functions for each of the G groups.
- For each hazard function, the baseline hazard function is different.
- The covariates $y_1, y_2, ..., y_k$ which violate the PH assumption are not included in the model.
- The parameter coefficients $\boldsymbol{\beta}$ are the same for each of the G groups.

A consequence of this model having a different baseline hazard function for each group is that the survival curves are different for each group, which can easily be seen in the derivation of the survival function for Cox's PH model (10).
Estimation of parameter coefficients $\boldsymbol{\beta}$ are obtained in a similar fashion to the way in which they are obtained for the original Cox PH model, as shown in section 3.3. However, for the SC PH model, the likelihood function L is the product of individual likelihoods of each group, $L_1, ..., L_G$, which are derived from their respective hazard functions $\lambda^g(t)$:

$$L = \prod_{g=1}^{G} L_g = \prod_{g=1}^{G} \frac{\lambda_0^g(t) \exp(\sum_{j=1}^{p} \beta_j x_j^g)}{\sum_{\alpha \in R^g(t_i)} \lambda_0^g(t) \exp(\sum_{j=1}^{p} \beta_j x_j^\alpha)}$$

Where $R^g(t_i)$ is the set of individuals with event time t_i and all individuals in group g at risk of experiencing the event time t_i.
The estimated parameter coefficients $\hat{\boldsymbol{\beta}}$ are then obtained in the same fashion as demonstrated in section 3.3.

6.2. **The No-interaction Assumption.**
As seen in the formula for the SC model (16), the parameter coefficients $\boldsymbol{\beta}$ are the same for each group of the variable Y. This is called the *no-interaction assumption*. From this assumption, we can see that the hazard ratio for each covariate is the same for each group and therefore in turn the covariates act the same on the baseline hazard $\lambda_0^g(t)$ for each group. This assumption can be tested by means of a Wald's test or likelihood ratio test (as described in sections 5.2 and 5.3 respectively). Below, we outline the details for performing the likelihood ratio test to check the no-interaction assumption.

Define a full interaction SC PH model as:

$$\lambda^g(t) = \lambda_0^g(t) \exp\left(\sum_{j=1}^{p} \beta_j^g x_j\right) \qquad \textit{(Note the coefficients depend on the group).} \tag{17}$$

Let $Y_1, ..., Y_{G-1}$ be defined as new dummy variables, obtained from the G groups of Y as defined in the prior subsection 6.1. Define a new, full interaction SC PH model as:

(18)
$$\lambda^g(t) = \lambda_0^g(t) \exp\left(\sum_{j=1}^{n} \beta_j x_j + \sum_{i=1}^{G-1} \sum_{j=1}^{n} \beta_j^i (Y_i \times x_j)\right) \quad \textit{(Involving dependent and independent coefficients)}$$

(17) and (18) can be proven to be equivalent[8, see page 183 for an example].
It can be seen that (17) is nested in (18) therefore, using the likelihood ratio test, we compare the maximised log-likelihood of the models involving the full interaction between PH violating covariates and PH satisfying covariates, and the model with no interaction between the two groups of covariates. The null hypothesis being that the no-interaction assumption holds, i.e.:

$$H_0 : \begin{cases} \beta_1^1 = ... = \beta_n^1 = 0 \\ \beta_1^2 = ... = \beta_n^2 = 0 \\ \vdots \\ \beta_1^{G-1} = ... = \beta_n^{G-1} = 0 \end{cases}$$

and alternative hypothesis that the no-interaction assumption does not hold:

$$H_A : \neg H_0.$$

The test statistic is given by

$$X^2_{LR} = 2[l_{\text{full}} - l_0]$$

where,

- l_{full} denotes the maximised log-likelihood of the model involving the full interaction between PH violating covariates and PH satisfying covariates;
- l_0 denotes the maximised log-likelihood of the model with no interaction between the two groups of covariates;

with

$$X^2_{LR} \overset{H_0}{\sim} \chi^2_{p(G-1)}.$$

We reject H_0 if for significance level α we find that $X^2_{LR} > X^2_{p(G-1)}(1-\alpha)$ where $X^2_{p(G-1)}(1-\alpha)$ denotes the $(1-\alpha)$ quantile for the Chi-squared distribution with $p(G-1)$ degrees of freedom, and conclude that the no-interaction assumption has been violated.

7. Application of Cox's PH Regression model to Peer-to-Peer Loan books

In this section, we apply the Survival Analysis methods discussed in prior sections to the real life data of peer-to-peer loans. The loan book data used in our analysis contains 108,904 individual loans commencing from April 2013 to December 2013, each with many characteristics given about borrower.
The grade of the loan is the primary characteristic of focus in this project, it is assigned to each loan by the Lending Club propriety scoring models. The propriety scoring model's score is then combined with the borrower's FICO score; credit attributes; and other application data to arrive at a base risk sub-grade. This method of assigning a grade to each loan is performed by internally developed algorithm (which is not publicly available). The base risk sub-grades assigned a model rank from 1-25 then modified using information such as the requested loan amount and the loan maturity (the planned time for which the loan is paid across), which formulates the final sub-grade 1-35. Sub-grades can then be categorised into alphabetical groups of 5, with A grades taking sub-graded loans 1-5; B grades 6-10; C grades 11-15; D grades 16-20; E grades 21-25; F grades 26-30; and finally G grades 31-35. A grades are considered the least risky of all categories, therefore inccuring the smallest interest rate on them (6.03% for A1 grades in our data), and G grades the most risky of all categories, therefore incurring the greatest interest rate (a massive 24.89% for G5 grades in our data)[23].
Often it is of interest to the lender to minimise the risk of investing in loans with a higher chance of defaulting, but also to maximise the profit gained by taking on riskier loans which have a greater interest rate. The loan book data in use for our analysis contains 13876 A grade loans (12.7%), 34946 B grade loans (32.1%), 30620 C grade loans (28.1%), 17172 D grade loans (15.8%), 7672 E grade loans (7%), 3820 F grade loans (3.5%) and 797 G grade loans (0.7%).

For the purpose of applying the Survival Analysis techniques discussed earlier, I have taken the survival times of each loan to be the difference in months between the loan commencing and the final payment date. All loans which were not defaulted or charged off were censored.

7.1. **Investigating the Influence of Grade on Probability of Loan Default.** In investigating the influence of the assigned grade of each loan to the probability of loan default, I investigated two different models: treating grade as both a linear numerical predictor (model 2) and treating each individual grade as a categorical predictor (model 1).

The model for treating each individual grade as a categorical predictor is:

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp\left(\beta_1 I_{X_1=A} + \beta_2 I_{X_2=B} + \beta_3 I_{X_3=C} + \beta_4 I_{X_4=D} + \beta_5 I_{X_5=E} + \beta_6 I_{X_6=F}\right)$$

where,

$$I_{X_j=\alpha} = \begin{cases} 1, & \text{if loan has grade } \alpha \\ 0, & \text{otherwise} \end{cases}$$

for all covariates j.

The reasoning behind not making an indicator variable for grade G is that it would create dependence between the 7 variables. This is because if you knew the values of each $I_{X_1=A}, ..., I_{X_6=F}$, then you would also know the value of $I_{X_7=G}$. The loans with grade G assigned to them are therefore said to be in the referent group, which could also be assigned to group A loans since the referent group is expected to be at either extreme of risk with all covariates being 0 [9, page 249]).

The model for treating the loan grade as a numerical predictor is

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(\beta A_X)$$

where,

$$A_X = \begin{cases} 1, & \text{if loan has grade A.} \\ 2, & \text{loan has grade B.} \\ 3, & \text{loan has grade C.} \\ 4, & \text{loan has grade D.} \\ 5, & \text{loan has grade E.} \\ 6, & \text{loan has grade F or G.} \end{cases}$$

I found that model 2 (the numerical predictor model) is highly probable to be inappropriate. To see this, let the hazard ratio of a loan with numerical grade i to a loan of grade with numerical value of $i-1$ denoted HR^i_{i-1} for all $i \in \{2,3,4,5,6\}$. All of these hazard ratios must equal each other under this model, i.e.

$$HR^i_{i-1} = \frac{\exp(6\eta)}{\exp(5\eta)} = \frac{\exp(5\eta)}{\exp(4\eta)} = ... = \frac{\exp(2\eta)}{\exp(\eta)} = \exp(\eta)$$

In words, this assumption states that the hazard risk of being in numerical group 6 in comparison to numerical group 5 is identical to the hazard risk of being in numerical group 5 in comparison to numerical group 4 and so forth. Such relationships between the grades are unlikely to be true, unless that is how the grades were chosen. We continue our analysis of loan default probability in relation to assigned grade using Model 1.

Figure 6 displays the estimated Cox PH regression survival curves for each grade A-G. As expected, the survival curves display a substantial less risk of default for grade A loans in comparison to grade G loans. The chance of default begins relatively similar in earlier months of the loans maturity, but as time passes the riskier nature of each

successive loan A-G becomes apparent as the survival curves diverge. The difference in chance of default between each successive grade appears to be relatively constant as time passes, except for grade G loans which appears to be only slightly less risky than grade F loans. This may be due to the small percentage of data being grade G loans (0.7%), and an even smaller percentage of those loans defaulting.

From Table 7:

(A) The mean survival time before default ranges from 54.3 months for A grade loans to 40.4 months for G grade loans. The median survival time is only available for grade G because it is the only grade which does not have more than 50% of its data censored. In order to estimate the median survival time for other grades, a parametric distribution such as the Weibull distribution could be applied and then the median could be extrapolated.

(B) Displays the estimated covariate coefficients and hazard ratios. Since G grade is taken as the referent group, all hazard ratios display the relative risk of default of that grade in comparison to grade G. The relative risk of default of grade A loans to grade G loans is seen to be 0.13, and 0.96 for F graded loans.

(C) Displays the correlation coefficient and corresponding p-values for the Schoenfield residuals for each covariate A-G. Since the p-values are small (≤ 0.05) for grades A and B it implies that these covariates do not satisfy the proportional hazards assumption.

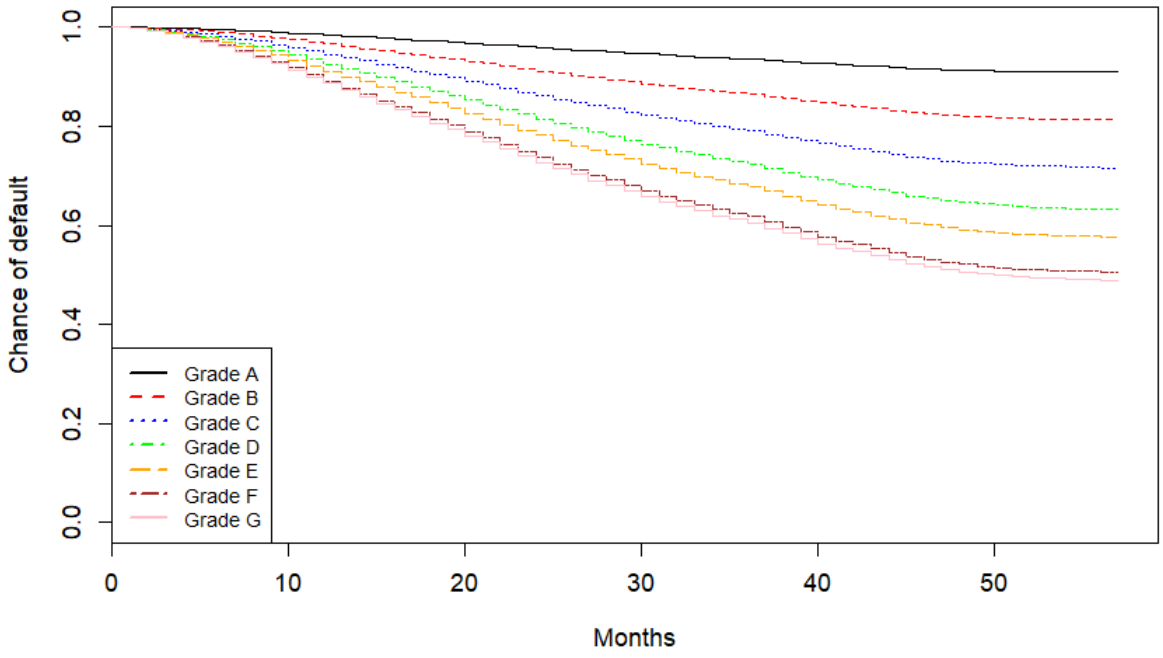


FIGURE 6. Estimated survival curves for each grade A-G under a Cox regression.

The Schoenfield residual plots seen in Figure 7 are not much help in determining any violations of the proportional hazards assumption here. The smooth red lines appear to be straight for all covariates A to G imply a zero correlation with the Schoenfield residuals and time, which contradicts the findings in table (6)(C). Further investigation by graphical techniques discussed previously is required. The comparison of Kaplan-Meier and Cox regression estimated survival curves in Figure 8 do not help our inquiry of which

(A)

Grade	μ_{t_k}	$x_{0.5}$
A	54.3	X
B	51.3	X
C	48.2	X
D	45.4	X
E	43.4	X
F	40.9	X
G	40.4	51.0

(B)

Grade	β_{grade}	$\exp(\beta_{grade})$
A	-2.01	0.13
B	-1.24	0.29
C	-0.76	0.47
D	-0.44	0.64
E	-0.26	0.77
F	-0.04	0.96
G	0	1

(C)

Grade	ρ	p-value
A	0.03104	6.00×10^{-5}
B	0.02535	1.04×10^{-3}
C	0.02238	0.382
D	0.01058	0.172
E	0.00999	0.197
F	0.00120	0.877
GLOBAL	NA	1.67×10^{-15}

TABLE 7. Mean, median, estimated coefficients, hazard ratios and Schoenfield residual tests for grades.

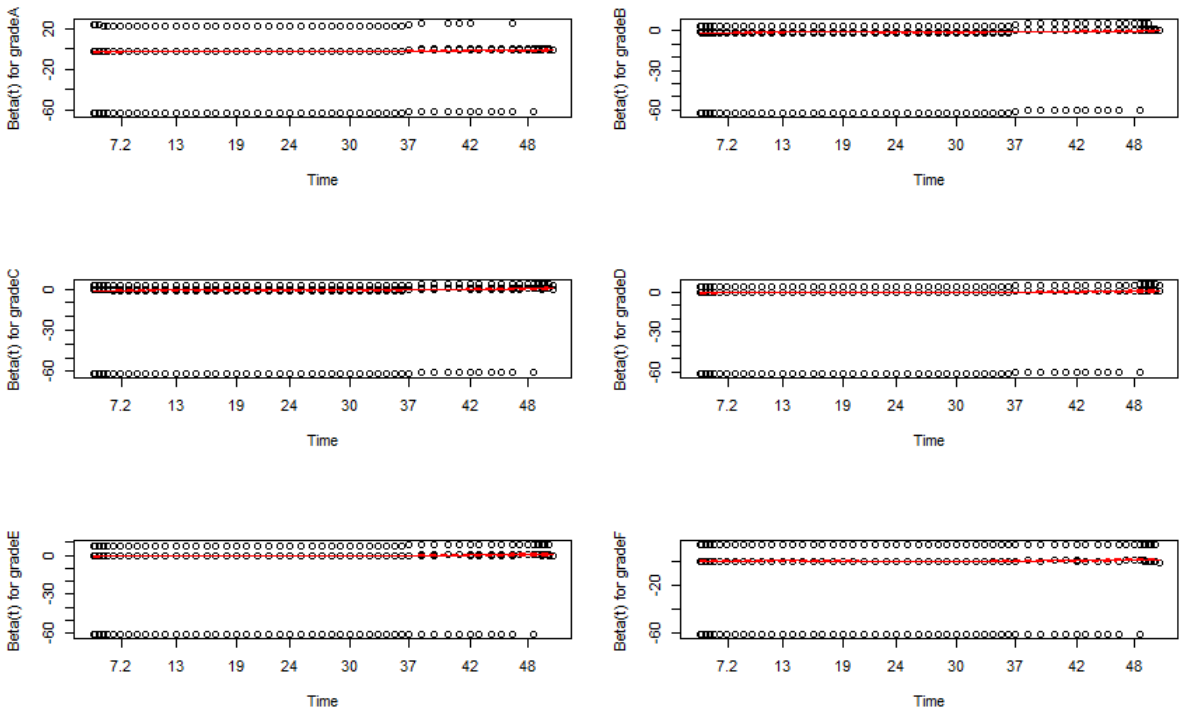


FIGURE 7. Scaled Schoenfield residual plots for grades A-F.

covariates satisfy the proportional hazards assumption. The survival curves for grades A to D seem to follow a similar pattern, implying they satisfy the proportional hazards assumption, and the survival curves for grades E to G seem to be distinctly different from one another, implying they do not satisfy the proportional hazards assumption.

The difference in log-log curves for as time passes for grades A-F appears to be relatively constant, implying that all covariates satisfy the proportional hazards assumption. The Schoenfield test for residuals, log-log survival curves and comparison of KM and Cox regression survival curve methods of determining whether covariates satisfy or violate the PH assumption all seem to contradict one another.

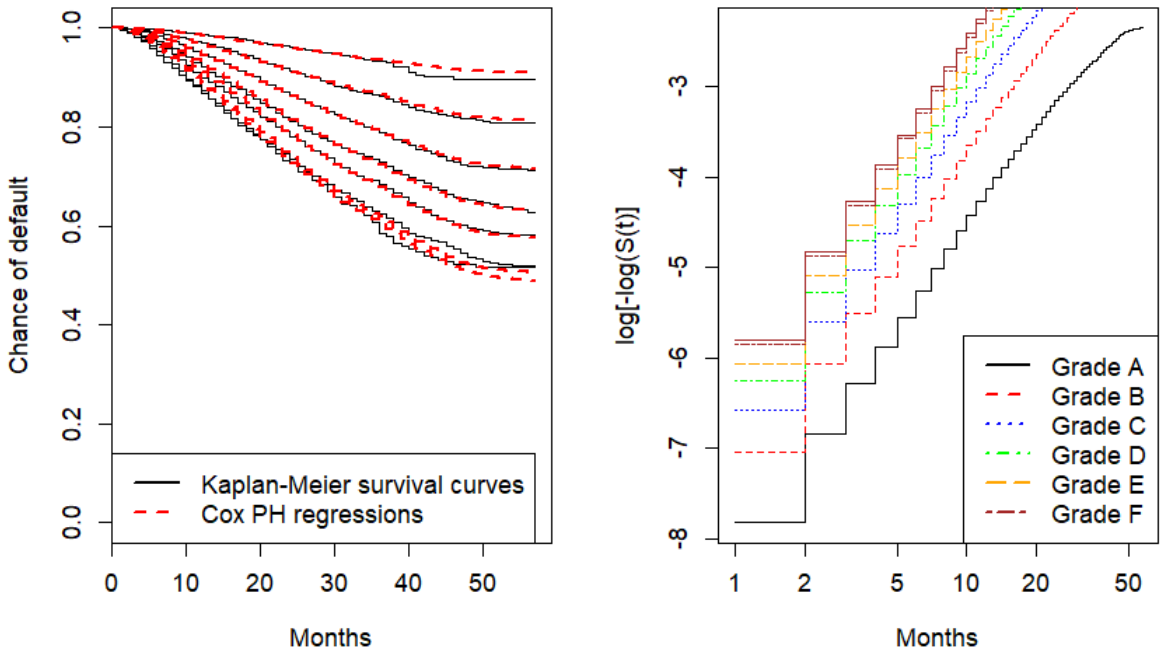


FIGURE 8. Comparison of Kaplan-Meier and Cox regression estimated survival curves and log-log survival curves for grades.

8. Conclusion, Summary of Results and Further Reading

This project has explored the required framework for a solid understanding on an introduction to Survival Analysis tailored to Cox's Proportional Hazard's model. The application of this mathematics to peer-to-peer loan book data was partially successful: the investigation of the appropriate model; plotted survival curves; hazard ratios; and mean survival times worked well providing some insight into the main goal of investigating the probability of default in relation to loan grade. However, the analysis came to a halt when all three discussed methods for determining violation of the proportional hazards assumption seemed to contradict one another. Some possible explanations for these odd results could be due to the large number of data or the large amount of censored data.
One possible route the reader could take for a further understanding on Survival Analysis is the topic of accelerated failure time models (AFT), a fully-parametric model which is an alternative to Cox's model. In addition to this, the reader could investigate into counting processes and their use in Survival Analysis. Fleming and Harrington dive into this topic in their excellent text *Counting Processes and Survival Analysis*[14].
This project was inspired by the research article *Determinants of Default in P2P lending (2015)* by Carlos Serrano-Cinca, Begona Gutierrez-Nieto and Luz Lopez-Palacios (C.B.P)[24]. C.B.P, unlike in my project, investigated the probability of default for loans due to many characteristics. My project hopefully demonstrated a suitable extension (with more of an emphasis on the mathematics behind the number) based on C.B.P's work, providing a unique analysis on the impact of a loan's assigned grade and the probability of default. Further work could investigate, using these results, potential investment strategies for maximising the return on investment whilst minimising the risk of investing in riskier loans.

References

[1] Bori, D. and Kavkler, A. (2009). Modeling Unemployment Duration in Slovenia using Cox Regression Models. Transition Studies Review: Volume 16, Issue 1. Pages 145-156. Retrieved from https://doi.org/10.1007/s11300-009-0053-6.

[2] Ihwah, A. (2015). The Use of Cox Regression Model to Analyze the Factors that Influence Consumer Purchase Decision on a Product. Agriculture and Agricultural Science Procedia: Volume 3, Pages 78-83. Retrieved from https://doi.org/10.1016/j.aaspro.2015.01.017.

[3] Ni, J. (2009). Application of Cox Proportional Hazard Model to the Stock Exchange Market. B.S. Undergraduate Mathematics Exchange: Volume 6, No. 1, pages 12-18. Retrieved from https://pdfs.semanticscholar.org/bd83/932aa17cecb385e0d156b1eda6cb359b70d5.pdf.

[4] Cox, D. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-220. Retrieved from http://www.jstor.org/stable/2985181

[5] Sir David R. Cox Wins Kettering Prize (1990). Chance journal: Volume 3, number 3, page 37. Publisher Taylor and Francis. Retrieved from https://doi.org/10.1080/09332480.1990.10554968.

[6] Retrieved from https://www.fundingcircle.com/uk/statistics/, Investor Returns. Accessed November 21 2017.

[7] Retrieved from https://www.zopa.com/about, Key Facts. Accessed November 21 2017.

[8] Klein, M and Kleinbaum, D.G. (2005). Survival Analysis: A Self-Learning Text, Third Edition. Published by Springer-Verlag New York. Retrieved from https://www.springer.com/gb/book/9781441966452.

[9] Klein, J.P and Moeschberger, M.L. (2003). Survival Analysis: Techniques for Censored and Truncated Data. Published by Springer-Verlag New York. Retrieved from https://www.springer.com/gb/book/9780387953991.

[10] Kaplan, E. L. and Meier.P. (1958). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association: Volume 53, No. 282, 1958, pages 457481. Retrieved from www.jstor.org/stable/2281868.

[11] Cox, D. (1975). Partial likelihood. Biometrika: Volume 62, Issue 2, 1 August 1975, Pages 269-276. Retrieved from http://www.jstor.org/stable/i315480.

[12] Hosmer D.W, Lemeshow S and May S. (2011). Applied Survival Analysis: Regression Modelling of Time to Event Data. Published by John Wiley and Sons Inc. Retrieved from https://onlinelibrary.wiley.com/doi/book/10.1002/9780470258019.

[13] Andersen P.K, Borgan , Gill R.D, Keiding N. (1993). Statistical Models Based on Counting Processes. Published by Springer-Verlag New York. Retrieved from https://www.springer.com/gb/book/9780387945194.

[14] Fleming T.R, Harrington D. (1991). Counting Processes and Survival Analysis. Published by John Wiley and Sons, Inc. Retrieved from https://onlinelibrary.wiley.com/doi/book/10.1002/9781118150672

[15] Breslow, N. Covariance Analysis of Censored Survival Data. (1974). Biometrics: Volume 30, No. 1, 1974, pages 8999. Retrieved from www.jstor.org/stable/2529620. Retrieved from https://doi.org/10.1093/biomet/62.2.269.

[16] Efron, B. (2012) .The Efficiency of Cox's Likelihood Function for Censored Data. Journal of the American Statistical Association: Volume 72, Issue 359, pages 557-565. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/01621459.1977.10480613.

[17] Prentice R.L and Kalbfleisch J.D. (2002). The Statistical Analysis of Failure Time Data, 2nd Edition. Pubished by John Wiley and Sons Inc. Retrieved from https://www.wiley.com/en-gb/The+Statistical+Analysis+of+Failure+Time+Data%2C+2nd+Edition-p-9780471363576.

[18] Hertz-Picciotto I, Rockhill B. (1997). Validity and efficiency of approximation methods for tied survival times in Cox regression: Biometrics, Volume 53, No.3, pages 11511156. Retrieved from https://www.jstor.org/stable/2533573.

[19] Harrell F.E, Califf R.M, Pryor D.B, Lee K.L and Rosati R.A. (1982). Evaluating the yield of medical tests. JAMA: Volume 246, No.18, pages 2463-2620. Retrieved from https://doi.org/10.1001/jama.1982.03320430047030.

[20] Schoenfield, D. (1982). Partial Residuals for the Rroportional Hazards Regression Model. Biometrika: Volume 69, Issue 1, pages 239 - 241. Retrieved from https://doi.org/10.1093/biomet/69.1.239.

[21] Cox, D, Hinkley, D. (1979). Theoretical Statistics. New York: Chapman and Hall/CRC. Retrieved from https://books.google.co.uk/books/about/Theoretical_Statistics.html?id=ppoujo-BInsC&redir_esc=y.
[22] P. Grambsch and T. Therneau. (1994). Proportional hazards tests and diagnostics based on weighted residuals. Biometrika: Volume 81, No. 3, pages 515-526. Retrieved from www.jstor.org/stable/2337123.
[23] https://www.lendingclub.com/foliofn/rateDetail.action. Accessed 09/04/18.
[24] Serrano-Cinca C, Gutirrez-Nieto B and Lpez-Palacios L. (2015). Determinants of Default in P2P Lending. Published by PLoS one. Retrieved from https://doi.org/10.1371/journal.pone.0139427.
[25] Retrieved from http://stat.ethz.ch/R-manual/R-devel/library/survival/html/Surv.html. Accessed April 22 2018.

R code for Figures and Tables

Below is the R code for all figures and data within tables from this project. Packages survival and survminer are used, for which more information can be obtained at https://cran.r-project.org/web/packages/survival/index.html (https://cran.r-project.org/web/packages/survival/index.html) and https://cran.r-project.org/web/packages/survminer/index.html (https://cran.r-project.org/web/packages/survminer/index.html) respectively. In order to run this code these packages must be installed, which can be done by entering `install.packages("survival")` and `install.packages("survminer")` into the R console. This document was composed using R notebooks. The .csv file containing the loan book data can be found at $https://drive.google.com/open?id=1\_8p53bvQigVb-awt2TlBZC8oxTd1iCe$. The tables and figures are listed in order of appearance.

Table 1: First 3 individuals of lung cancer time-to-event data.

```
library(survival)
head(lung)
```

	inst <dbl>	time <dbl>	status <dbl>	age <dbl>	sex <dbl>	ph.ecog <dbl>	ph.karno <dbl>	pat.karno <dbl>	meal.cal <dbl>
1	3	306	2	74	1	1	90	100	1175
2	3	455	2	68	1	0	90	90	1225
3	3	1010	1	56	1	0	90	90	*NA*
4	5	210	2	57	1	1	90	60	1150
5	1	883	2	60	1	0	100	90	*NA*
6	12	1022	1	74	1	1	50	80	513

6 rows | 1-10 of 11 columns

Figure 1: Survival curves for male and female patients with lung cancer. These curves are fitted using the Kaplan-Meier (KM) estimate for the survival function $\hat{S}(t)$. Note the step-wise nature of the plot, indicating the constant survival probabilites for all time between successive survival times. Dotted lines indicate confidence 95% confidence intervals.

```
library(survival) #loads survival package used for survival analysis techniques.
library(survminer) #loads survminer package used for fancy survival curve graphics.
```

```
## Warning: package 'survminer' was built under R version 3.4.4
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggpubr
```

```
## Warning: package 'ggpubr' was built under R version 3.4.4
```

```
## Loading required package: magrittr
```

```
lung1<-na.omit(lung) #removes all data sets with an observed result of NA. Very few entries h
ad incomplete data, which malfunctions code for computing Schoenfield residuals later on if n
ot removed.
model<-survfit(Surv(time,status)~sex,data=lung1) #estimates survival function using kaplan-me
ier estimate
ggsurvplot(model,risk.table=TRUE,risk.table.fontsize=3,conf.int=TRUE,linetype="strata",legen
d.labs=c("Male","Female"))#creates survival curve and risk table.
```

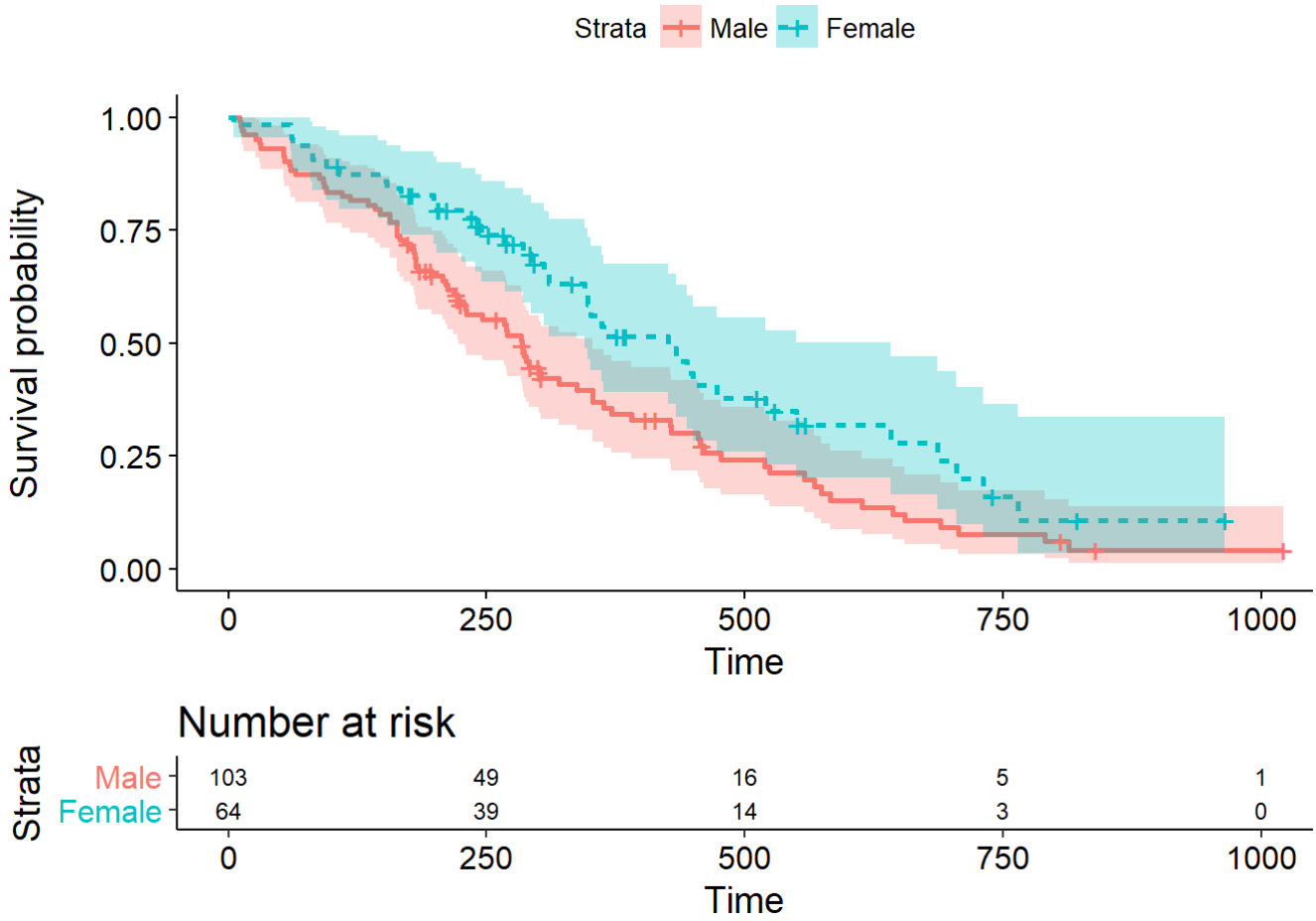


Table 2: Displays the values of t_i, r_i, d_i, $S(t_i)$ and $Var(\hat{S}(t_i))$ and the 95 percent confidence interval for the first 3 event times. These agree with the results obtained below, for which time=t_i, n.risk=r_i, n.event=d_i, survival=$S(t_i)$ and std.err=$e.s.e(\hat{S}(t_i))$.

```
library(survival)
lung1<-na.omit(lung) #removes all data sets with an observed result of NA.
model<-survfit(Surv(time,status)~sex,data=lung1) #Estimates survival function for lung cancer
 data with covariate sex by means of the KM method.
modelout<-capture.output(summary(model))#prints out first 8 lines of function summary of mode
l fitted.
head(modelout,8)
```

```
## [1] "Call: survfit(formula = Surv(time, status) ~ sex, data = lung1)"
## [2] ""
## [3] "                sex=1 "
## [4] " time n.risk n.event survival std.err lower 95% CI upper 95% CI"
## [5] "   11    103       1   0.9903 0.00966       0.9715        1.000"
## [6] "   12    102       1   0.9806 0.01360       0.9543        1.000"
## [7] "   13    101       1   0.9709 0.01657       0.9389        1.000"
## [8] "   15    100       1   0.9612 0.01904       0.9246        0.999"
```

Table 3: Estimates μ_{t_k} & $\hat{x}_{0.5}$ for male and female patients with lung cancer.

```
library(survival)
lung1<-na.omit(lung) #removes all data sets with an observed result of NA.
model<-coxph(Surv(time,status)~sex,data=lung1)#fits Cox's model to the data.
print(survfit(model,newdata=data.frame(sex=1)),rmean="common")#outputs data quantities for ma
le patients
```

```
## Call: survfit(formula = model, newdata = data.frame(sex = 1))
##
##          n     events     *rmean *se(rmean)     median    0.95LCL
##      167.0      120.0      341.8       18.3      286.0      245.0
##    0.95UCL
##      353.0
##     * restricted mean with upper limit =  1022
```

```
print(survfit(model,newdata=data.frame(sex=2)),rmean="common")#outputs data quantities for fe
male patients
```

```
## Call: survfit(formula = model, newdata = data.frame(sex = 2))
##
##          n     events     *rmean *se(rmean)     median    0.95LCL
##      167.0      120.0      465.5       35.2      426.0      320.0
##    0.95UCL
##      567.0
##     * restricted mean with upper limit =  1022
```

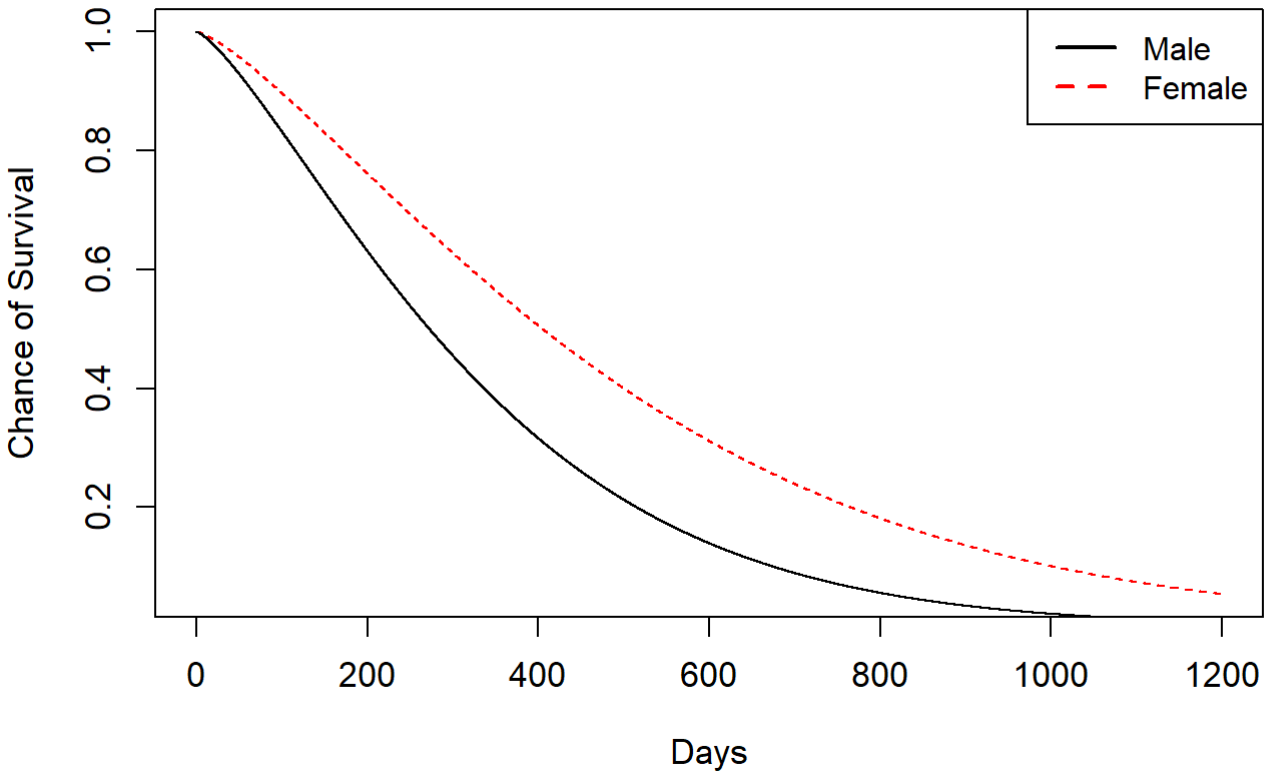
Table 4: The estimated coefficients and hazard ratios for the model fitting the covariates sex and age.The covariate age is seen to have a slight positive association with the event hazard, meaning that older patients have a higher chance of dying. The covariate sex is a dichotomous variable, with 1 representing males and 2 representing females, therefore we can conclude that males have a higher chance of dying due to advanced lung cancer.

```
library(survival)
lung1<-na.omit(lung) #removes all data sets with an observed result of NA
model<-coxph(Surv(time,status)~sex+age,data=lung1)#fits Cox's model to the data.
modelout1<-capture.output(summary(model))
head(modelout1,8)
```

```
## [1] "Call:"
## [2] "coxph(formula = Surv(time, status) ~ sex + age, data = lung1)"
## [3] ""
## [4] "  n= 167, number of events= 120 "
## [5] ""
## [6] "        coef exp(coef) se(coef)      z Pr(>|z|)  "
## [7] "sex -0.44669   0.63974  0.19754 -2.261   0.0237 *"
## [8] "age  0.01737   1.01752  0.01084  1.603   0.1089  "
```

Figure 2: Comparison of survival curves. First figure : Weibull AFT applied to lung data. Second figure : Cox's Proportional Hazards model applied to same lung data. The Weibull AFT parametric model is seen to be very similar to the plotted Cox PH regression model, which demonstrates the robust nature of Cox's PH model.

```
library(survival)
lung1<-na.omit(lung) #removes all data sets with an observed result of NA
model1<-survreg(Surv(time,status)~sex,data=lung1,dist="weibull")
alpha<-1/0.755
beta.m<-exp(5.885)
beta.f<-exp(6.281)
time<-seq(0,1200,by=.1)
surv.f<-1-pweibull(time,alpha,beta.f)
surv.m<-1-pweibull(time,alpha,beta.m)
plot(time,surv.f,col="red",type="l",lwd=1,ylab="Chance of Survival",xlab="Days",lty=2,cex.axi
s=1.1,cex.lab=1.1)#plots weibull survival curve for females
lines(time,surv.m,lwd=1) #Adds male weibull survival curve to plot
legend(x="topright",legend=c("Male","Female"),col=c("black","red"),lty=c(1,2),lwd=2)
```



```
model<-coxph(Surv(time,status)~sex,data=lung1)
plot(survfit(model,newdata=data.frame(sex=1),conf.type="none"),xlab="Days",ylab="Chance of Su
rvival",lwd=1,lty=1,cex.axis=1.1,cex.lab=1.1)
lines(survfit(model,newdata=data.frame(sex=2),conf.type="none"),xlab="Days",ylab="Chance of S
urvival",lwd=1,col="red",lty=2)
legend(x="topright",legend=c("Male","Female"),col=c("black","red"),lty=c(1,2),lwd=2)
```

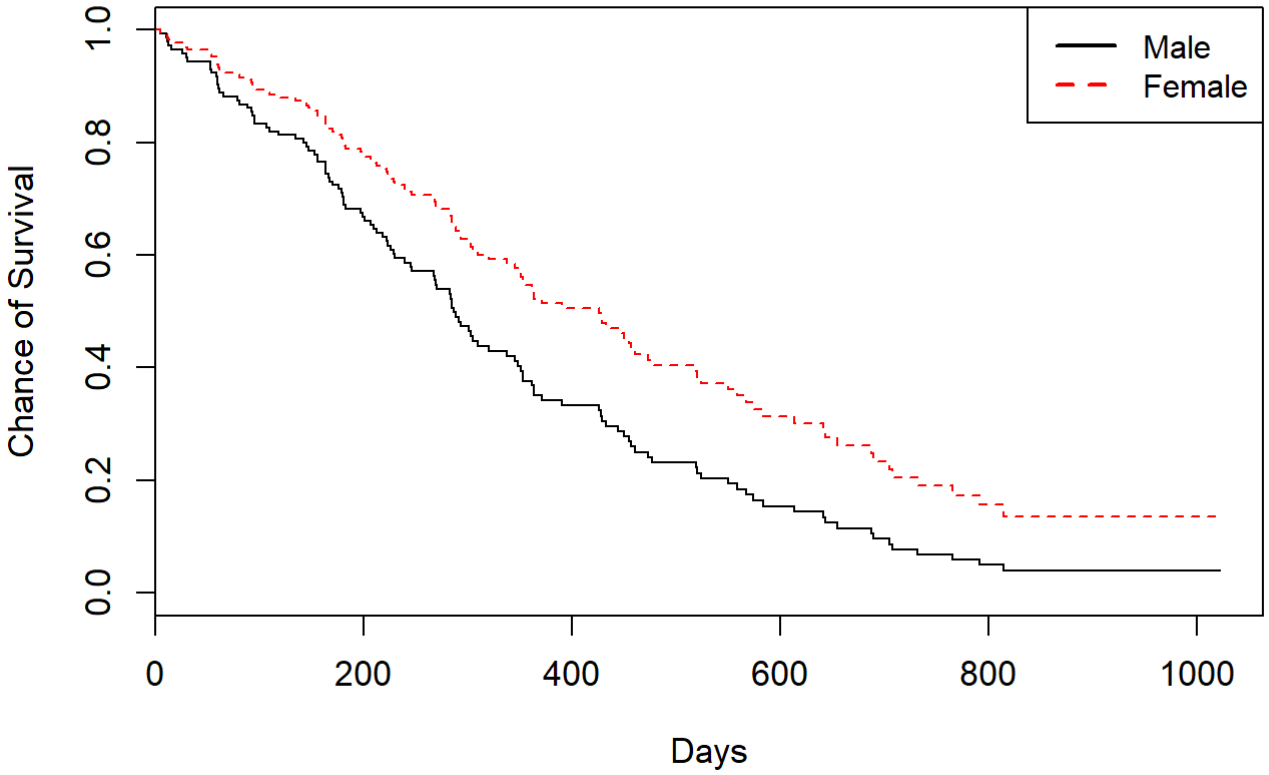


Figure 3: $\log(-\log(S(t))$ plots of survival curves for male and females. The distance between each curve seems to be relatively constant throughout the plot,with a difference of $\sum_{j=1}^{p} \beta_j(x_j^a - x_j^b) \approx \sum_{j=1}^{1} -0.447(1-2) \approx 0.447$, therefore we can conclude that the covariate sex satisfies the proportional hazards assumption.

```
library(survival)
lung1<-na.omit(lung) #removes all data sets with an observed result of NA
model1<-coxph(Surv(time,status)~sex,data=lung1)
plot(survfit(model1,newdata=data.frame(sex=1),conf.type="none"),fun="cloglog",xlab="Days",yla
b="log[-log(S(t)]",lwd=1,cex.axis=1.1,cex.lab=1.1)
legend(x="topleft",legend=c("Male","Female"),col=c("black","red"),lty=c(1,2),lwd=2,cex=0.75)
lines(survfit(model1,newdata=data.frame(sex=2),conf.type="none"),fun="cloglog",col="red",lwd=
1,lty=2)
```

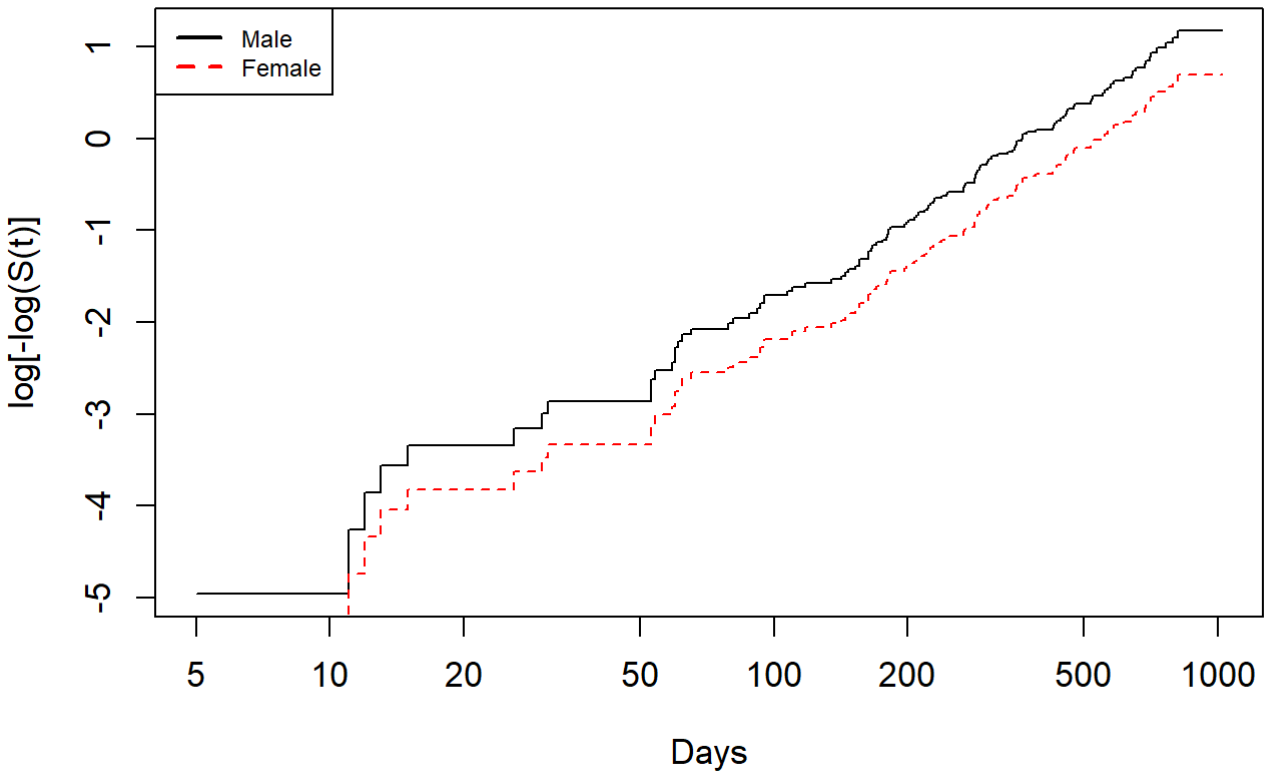


Figure 4: Comparison of estimated survival curves between Kaplan-Meier estimates and Cox regression for covariate sex in lung data. Plots show survival curves for males and females. We can see that the two different survival curves for males and females are relatively similar, therefore we can conclude that the covariate sex satisfies the proportional hazards assumption.

```
library(survival)
lung1<-na.omit(lung) #removes all data sets with an observed result of NA
plot(survfit(Surv(time,status)~sex,data=lung1),xlab="Days",ylab="Chance of Survival",lwd=1,ce
x.axis=1.1,cex.lab=1.1)
model1<-coxph(Surv(time,status)~sex,data=lung1)
lines(survfit(model1,newdata=data.frame(sex=2),conf.type="none"),col="Red",lwd=2,lty=2)
lines(survfit(model1,newdata=data.frame(sex=1),conf.type="none"),col="Red",lwd=2,lty=2)
legend(x="topright", legend=c("Kaplan-Meier estimated survival curve", "Cox PH estimated surv
ival curve"), col=c("black", "red"), lty=c(1,2), lwd=2, cex=1.1)
```

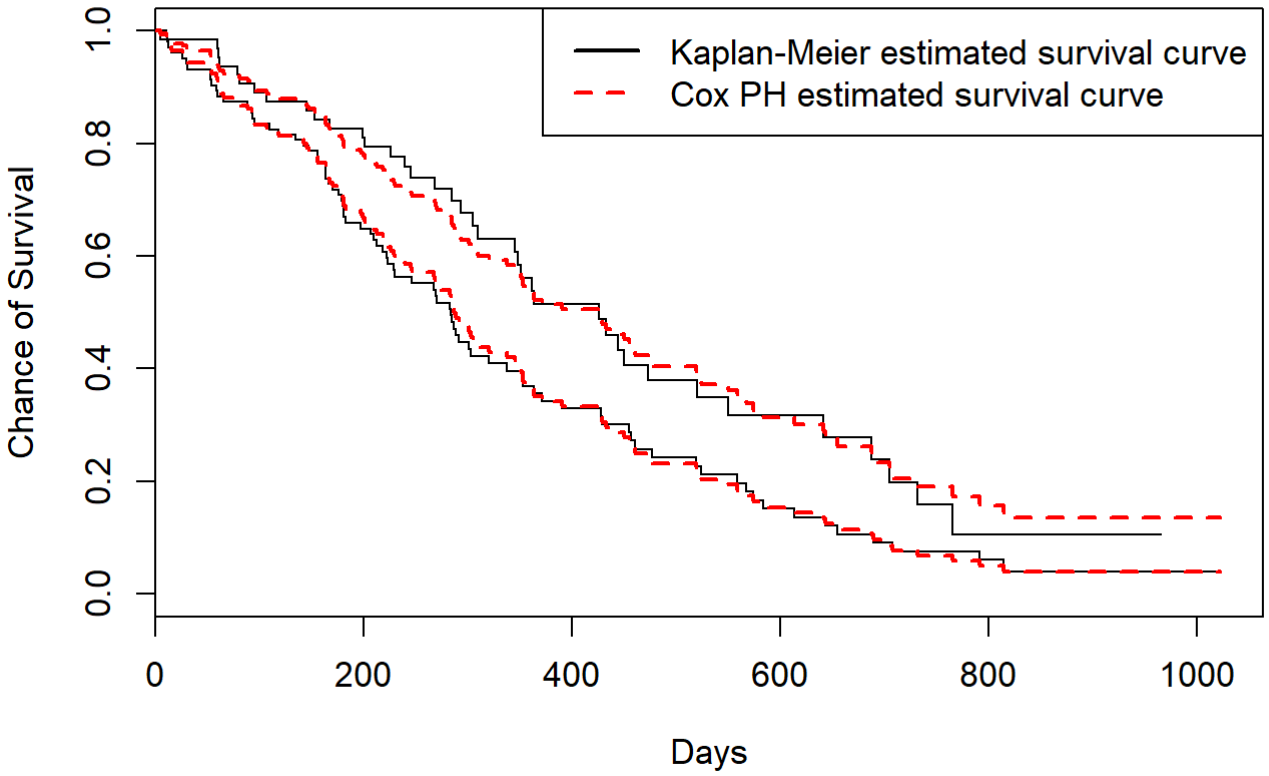
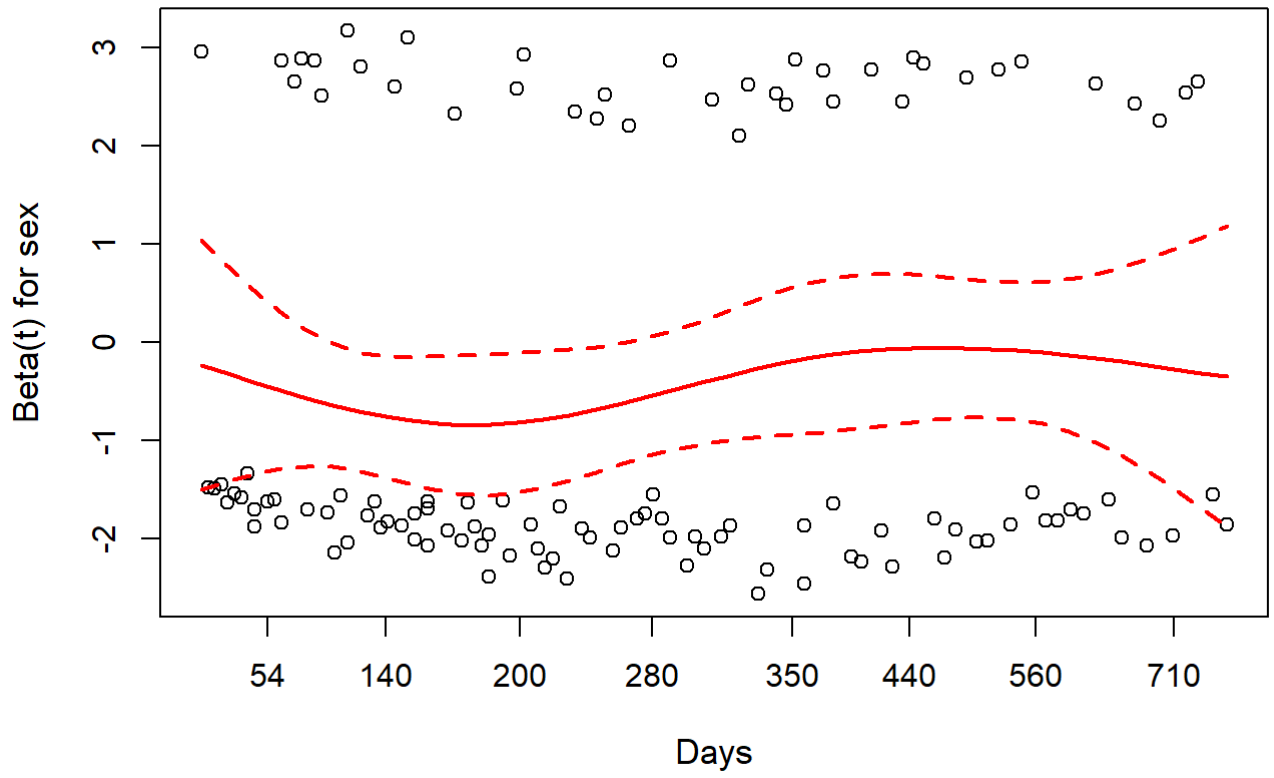


Figure 5 and Table 5: Scaled Schoenfield residuals plotted against time for covariates sex (left) and age (right). Accept $H_0 : \rho = 0$ for both covariates since the p-values for each covariate is not small and the smooth red lines are relatively straight. The correlation coefficient is 0 and the Schoenfield residuals do not vary with time and therefore the proportional hazards assumption is satisfied.

```
library(survival)
lung1<-na.omit(lung) #removes all data sets with an observed result of NA
model3<-coxph(Surv(time,status)~sex+age,data=lung1)
print(cox.zph(model3))
```

```
##            rho chisq     p
## sex     0.0841 0.816 0.366
## age    -0.0633 0.491 0.483
## GLOBAL      NA 1.444 0.486
```

```
plot(cox.zph(model3),lwd=2,col="red",xlab="Days",cex.axis=1.1,cex.lab=1.1)
```



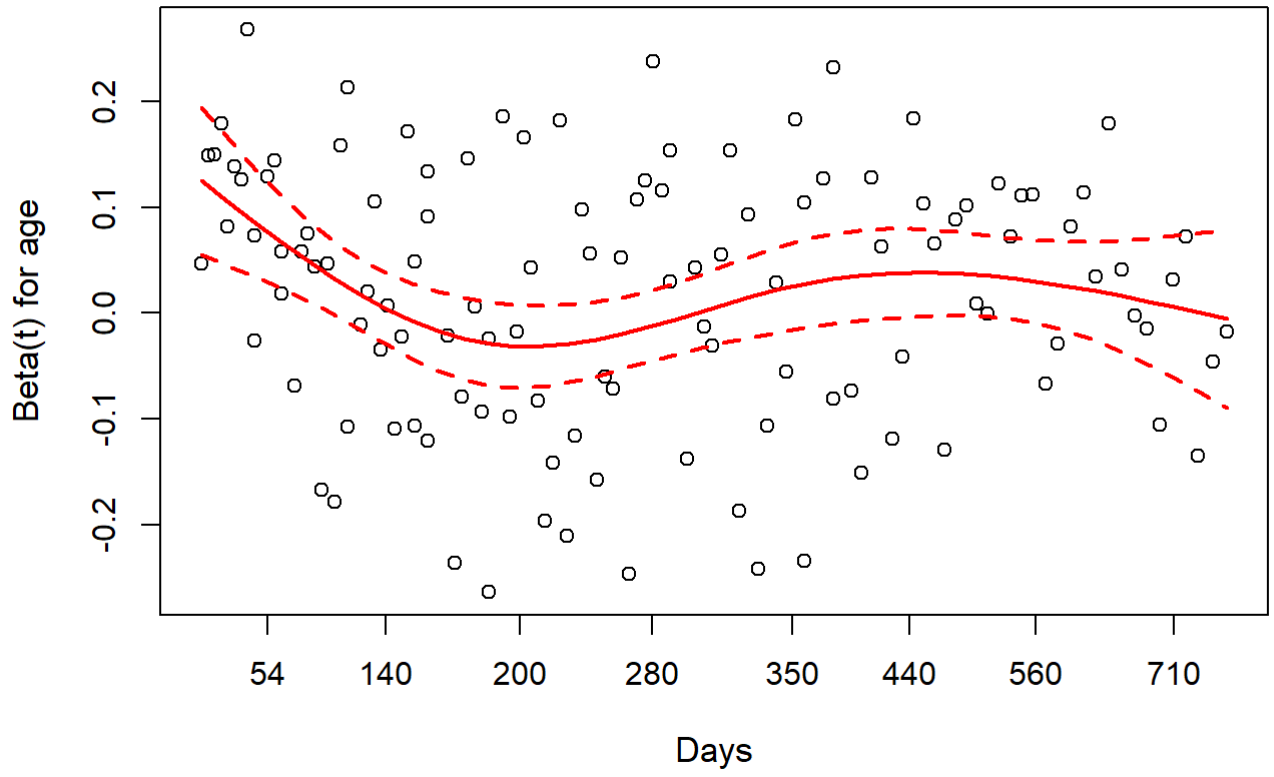


Table 6: Test statistic values for likelihood ratio, Wald test and score test. Since all test statistics are greater than $X_2^2(0.95)$ we can reject the null hypothesis.

```
library(survival)
lung1<-na.omit(lung) #removes all data sets with an observed result of NA
model<-coxph(Surv(time,status)~sex+age,data=lung1)
modelout1<-capture.output(summary(model))
tail(modelout1,4)
```

```
## [1] "Likelihood ratio test= 8.88  on 2 df,   p=0.01179"
## [2] "Wald test            = 8.47  on 2 df,   p=0.01449"
## [3] "Score (logrank) test = 8.6  on 2 df,   p=0.0136"
## [4] ""
```

Figure 6: Estimated survival curves for each grade A-G under a Cox regression model.

```
library(survival)
loandata<-read.csv(file.choose(),header=T) #This line of code will prompt the user to select
 the loan book data file, which must be selected before any further code is ran.
loandata$grade<-relevel(loandata$ï..grade,ref="G")
model.grade<-coxph(Surv(time,status)~grade,data=loandata)
plot(survfit(model.grade,newdata=data.frame(grade="A"),conf.type="none"),lwd=1,xlab="Months",
ylab="Chance of default",cex.axis=1.1,cex.lab=1.1)
lines(survfit(model.grade,newdata=data.frame(grade="B"),conf.type="none"),col="red",lwd=1,xla
b="Months",ylab="Chance of default",lty=2)
lines(survfit(model.grade,newdata=data.frame(grade="C"),conf.type="none"),col="blue",lwd=1,xl
ab="Months",ylab="Chance of default",lty=3)
lines(survfit(model.grade,newdata=data.frame(grade="D"),conf.type="none"),col="green",lwd=1,x
lab="Months",ylab="Chanc# of default",lty=4)
lines(survfit(model.grade,newdata=data.frame(grade="E"),conf.type="none"),col="orange",lwd=1,
xlab="Months",ylab="Chan#ce of default",lty=5)
lines(survfit(model.grade,newdata=data.frame(grade="F"),conf.type="none"),col="brown",lwd=1,x
lab="Months",ylab="Chance of default",lty=6)
lines(survfit(model.grade,newdata=data.frame(grade="G"),conf.type="none"),col="pink",lwd=1,xl
ab="Months",ylab="Chance of default")
legend(x="bottomleft", legend=c("Grade A", "Grade B","Grade C","Grade D","Grade E","Grade F",
"Grade G"), col=c("black", "red","blue","green","orange","brown","pink"), lty=c(1,2,3,4,5,6,1
), lwd=2, cex=0.9)
```

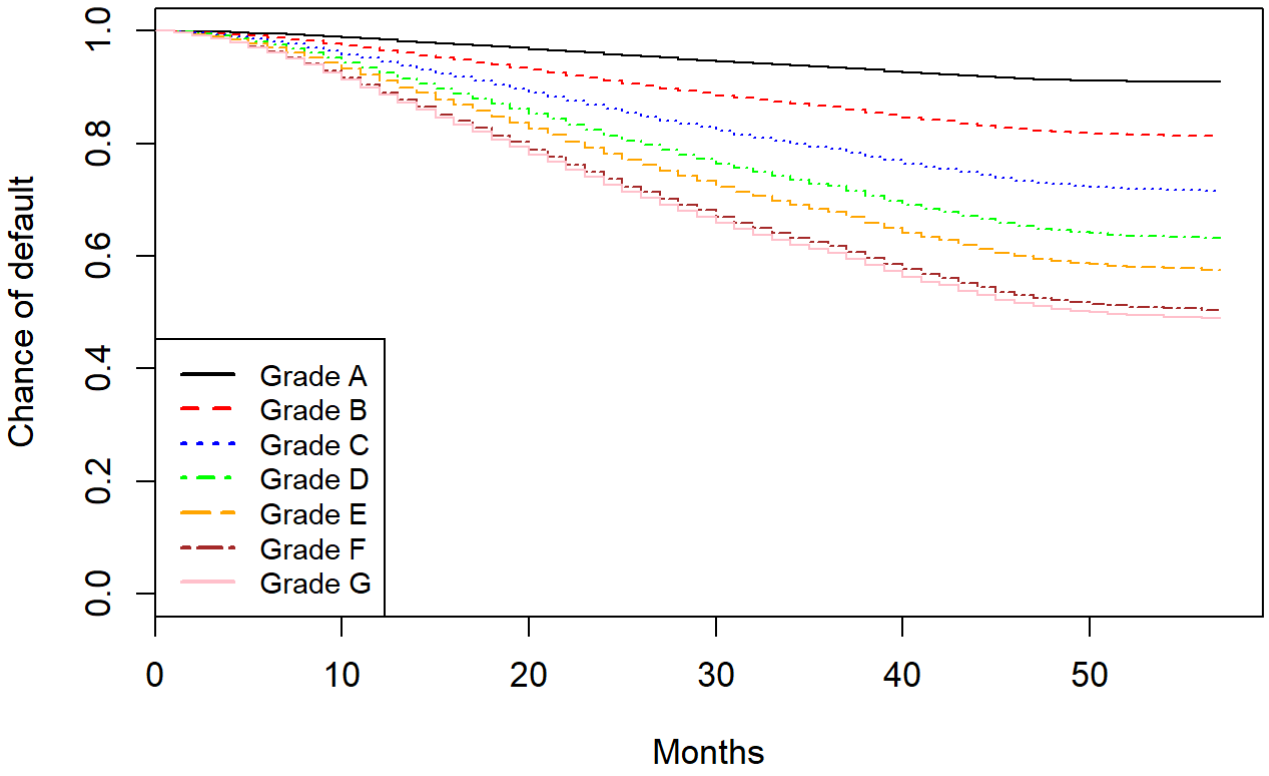


Figure 6: Estimated survival curves for each grade A-G under a Cox regression model.

```
print(survfit(model.grade,newdata=data.frame(grade="A")),rmean="common")
```

```
## Call: survfit(formula = model.grade, newdata = data.frame(grade = "A"))
##
##          n     events      *rmean *se(rmean)     median    0.95LCL
##   1.09e+05   1.67e+04   5.43e+01   7.38e-02         NA         NA
##    0.95UCL
##         NA
##     * restricted mean with upper limit =  57
```

```
print(survfit(model.grade,newdata=data.frame(grade="B")),rmean="common")
```

```
## Call: survfit(formula = model.grade, newdata = data.frame(grade = "B"))
##
##          n     events      *rmean *se(rmean)     median    0.95LCL
##   1.09e+05   1.67e+04   5.13e+01   6.76e-02         NA         NA
##    0.95UCL
##         NA
##     * restricted mean with upper limit =  57
```

```
print(survfit(model.grade,newdata=data.frame(grade="C")),rmean="common")
```

```
## Call: survfit(formula = model.grade, newdata = data.frame(grade = "C"))
##
##          n     events     *rmean *se(rmean)     median    0.95LCL
##   1.09e+05   1.67e+04   4.82e+01   6.13e-02         NA         NA
##    0.95UCL
##         NA
##     * restricted mean with upper limit =  57
```

```
print(survfit(model.grade,newdata=data.frame(grade="D")),rmean="common")
```

```
## Call: survfit(formula = model.grade, newdata = data.frame(grade = "D"))
##
##          n     events     *rmean *se(rmean)     median    0.95LCL
##   1.09e+05   1.67e+04   4.54e+01   5.57e-02         NA         NA
##    0.95UCL
##         NA
##     * restricted mean with upper limit =  57
```

```
print(survfit(model.grade,newdata=data.frame(grade="E")),rmean="common")
```

```
## Call: survfit(formula = model.grade, newdata = data.frame(grade = "E"))
##
##          n     events     *rmean *se(rmean)     median    0.95LCL
##   1.09e+05   1.67e+04   4.34e+01   5.19e-02         NA         NA
##    0.95UCL
##         NA
##     * restricted mean with upper limit =  57
```

```
print(survfit(model.grade,newdata=data.frame(grade="F")),rmean="common")
```

```
## Call: survfit(formula = model.grade, newdata = data.frame(grade = "F"))
##
##          n     events     *rmean *se(rmean)     median    0.95LCL
##   1.09e+05   1.67e+04   4.09e+01   4.71e-02         NA   4.90e+01
##    0.95UCL
##         NA
##     * restricted mean with upper limit =  57
```

```
print(survfit(model.grade,newdata=data.frame(grade="G")),rmean="common")
```

```
## Call: survfit(formula = model.grade, newdata = data.frame(grade = "G"))
##
##          n     events     *rmean *se(rmean)     median    0.95LCL
##   1.09e+05   1.67e+04   4.04e+01   4.60e-02   5.10e+01   4.40e+01
##    0.95UCL
##         NA
##     * restricted mean with upper limit =  57
```

```
summary(model.grade)$coef
```

```
##               coef exp(coef)   se(coef)           z     Pr(>|z|)
## gradeA -2.01472294 0.1333573 0.07213842 -27.9285690 0.000000e+00
## gradeB -1.23755185 0.2900935 0.06293805 -19.6630147 0.000000e+00
## gradeC -0.76065752 0.4673590 0.06212931 -12.2431341 0.000000e+00
## gradeD -0.44151785 0.6430596 0.06276200  -7.0347955 1.995515e-12
## gradeE -0.25753753 0.7729526 0.06437077  -4.0008457 6.311650e-05
## gradeF -0.04280795 0.9580954 0.06674479  -0.6413677 5.212838e-01
```

```
print(cox.zph(model.grade))
```

```
##            rho  chisq        p
## gradeA 0.03104 16.104 6.00e-05
## gradeB 0.02535 10.747 1.04e-03
## gradeC 0.02238  8.367 3.82e-03
## gradeD 0.01058  1.869 1.72e-01
## gradeE 0.00999  1.666 1.97e-01
## gradeF 0.00120  0.024 8.77e-01
## GLOBAL      NA 81.544 1.67e-15
```

Figure 7: Scaled Schoenfield residual plots for grades A-F

```
par(mfrow=c(3,2))
plot(cox.zph(model.grade),col="red",cex.lab=1.1,cex.axis=1.1)
```

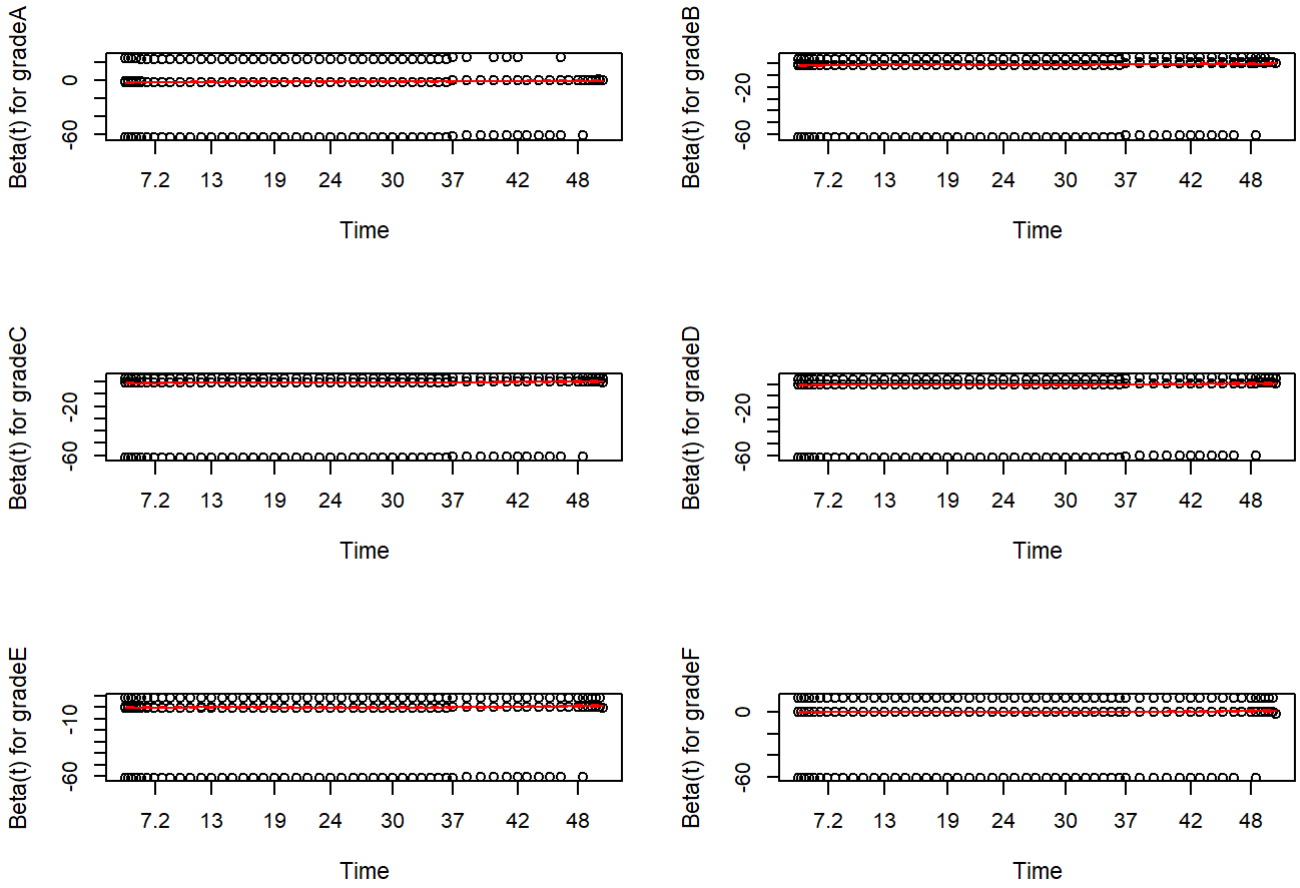
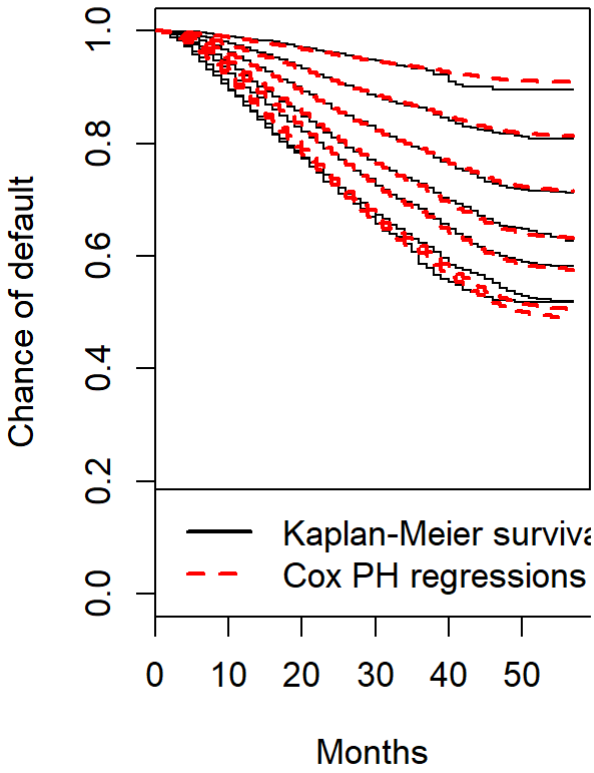
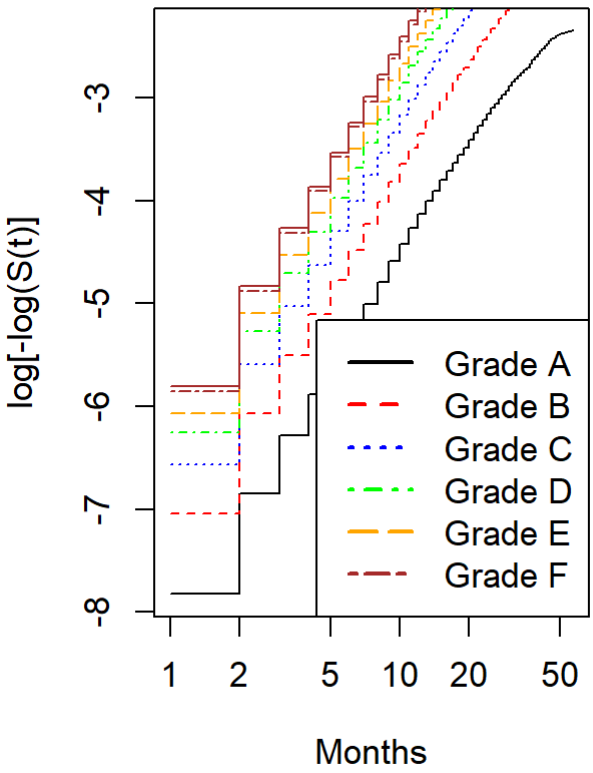


Figure 8: Comparison of Kaplan-Meier and Cox regression estimated survival curves and log-log survival curves for grades

```
par(mfrow=c(1,2))
plot(survfit(Surv(time,status)~ï..grade,data=loandata),xlab="Months",ylab="Chance of default"
,lty=1,cex.lab=1.1,cex.axis=1.1)
lines(survfit(model.grade,newdata=data.frame(grade="A"),conf.type="none"),col="red",lwd=2,xla
b="Months",ylab="Chance of default",lty=2)
lines(survfit(model.grade,newdata=data.frame(grade="B"),conf.type="none"),col="red",lwd=2,xla
b="Months",ylab="Chance of default",lty=2)
lines(survfit(model.grade,newdata=data.frame(grade="C"),conf.type="none"),col="red",lwd=2,xla
b="Months",ylab="Chance of default",lty=2)
lines(survfit(model.grade,newdata=data.frame(grade="D"),conf.type="none"),col="red",lwd=2,xla
b="Months",ylab="Chance of default",lty=2)
lines(survfit(model.grade,newdata=data.frame(grade="E"),conf.type="none"),col="red",lwd=2,xla
b="Months",ylab="Chance of default",lty=2)
lines(survfit(model.grade,newdata=data.frame(grade="F"),conf.type="none"),col="red",lwd=2,xla
b="Months",ylab="Chance of default",lty=2)
lines(survfit(model.grade,newdata=data.frame(grade="G"),conf.type="none"),col="red",lwd=2,xla
b="Months",ylab="Chance of default",lty=2)
legend(x="bottomleft", legend=c("Kaplan-Meier survival curves", "Cox PH regressions"), col=c(
"black", "red"), lty=c(1,2), lwd=2, cex=1.1)
plot(survfit(model.grade,newdata=data.frame(grade="A"),conf.type="none"),fun="cloglog",xlab=
"Months",ylab="log[-log(S(t)]",lwd=1,,cex.lab=1.1,cex.axis=1.1)
lines(survfit(model.grade,newdata=data.frame(grade="B"),conf.type="none"),fun="cloglog",xlab=
"Months",ylab="log[-log(S(t)]",lwd=1,col="red",lty=2)
lines(survfit(model.grade,newdata=data.frame(grade="C"),conf.type="none"),fun="cloglog",xlab=
"Months",ylab="log[-log(S(t)]",lwd=1,col="blue",lty=3)
lines(survfit(model.grade,newdata=data.frame(grade="D"),conf.type="none"),fun="cloglog",xlab=
"Months",ylab="log[-log(S(t)]",lwd=1,col="green",lty=4)
lines(survfit(model.grade,newdata=data.frame(grade="E"),conf.type="none"),fun="cloglog",xlab=
"Months",ylab="log[-log(S(t)]",lwd=1,col="orange",lty=5)
lines(survfit(model.grade,newdata=data.frame(grade="F"),conf.type="none"),fun="cloglog",xlab=
"Months",ylab="log[-log(S(t)]",lwd=1,col="brown",lty=6)
lines(survfit(model.grade,newdata=data.frame(grade="G"),conf.type="none"),fun="cloglog",xlab=
"Months",ylab="log[-log(S(t)]",lwd=1,col="brown",lty=1)
legend(x="bottomright", legend=c("Grade A", "Grade B","Grade C","Grade D","Grade E","Grade F"
), col=c("black", "red","blue","green","orange","brown"), lty=c(1,2,3,4,5,6,1), lwd=2,cex=1.1
)
```


Chance of default
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Months
Kaplan-Meier surviva
Cox PH regressions


log[-log(S(t))]
-8
-7
-6
-5
-4
-3
1
2
5
10
20
50
Months
Grade A
Grade B
Grade C
Grade D
Grade E
Grade F